On mass spectrometry analysis of gene regulatory protein-DNA interactions

Matthew Makowski

On mass spectrometry analysis of gene regulatory protein-DNA interactions

Matthew Makowski

The research described in this thesis was carried out at the Department of Molecular Biology, Faculty of Science, Radboud University Nijmegen, the Netherlands in conjunction with the Radboud Institute for Molecular Life Sciences. Financial support was provided by an FP7 grant for the Marie Curie Initial Training Network (ITN) DevCom (Main Applicant Gert Jan Veenstra). Additional funding granted to Michiel Vermeulen through the Oncode Institute which is partly financed by the Dutch Cancer Society (KWF) and was funded by the gravitation program CancerGenomiCs.nl from the Netherlands Organization for Scientific Research (NWO).

Lay-out Nikki Vermeulen | Ridderprint BV
Printing Ridderprint BV | www.ridderprint.nl

On mass spectrometry analysis of gene regulatory protein-DNA interactions

Proefschrift
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op

Vrijdag, 6 Juli 2018

om 14.30 uur precies

door

Matthew Michael Makowski
geboren op 14 November 1989
te Grand Forks, North Dakota, the United States of America

Promoter:
prof. dr. Michiel Vermeulen

Manuscriptcommissie:
prof. dr. Gert Jan Veenstra (Voorzitter)
prof. dr. Peter Friedl
dr. Tieneke Lenstra (Nederlands Kanker Instituut)

Paranimfen
Ino Karemaker
Christopher Makowski

On mass spectrometry analysis of gene regulatory protein-DNA interactions

Doctoral thesis
to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on

Friday, 6 July 2018

at 14.30 hours

by

Matthew Michael Makowski
born on 14 November, 1989
in Grand Forks, North Dakota, the United States of America

Supervised by:
prof. dr. Michiel Vermeulen

Members of the Manuscript Committee
prof. dr. Gert Jan Veenstra (Chair)
prof. dr. Peter Friedl
dr. Tieneke Lenstra (Netherlands Cancer Institute)

Paranymphs
Ino Karemaker
Christopher Makowski

Contents

Chapter 1	Introduction	9
Chapter 2	Recurrent promoter mutations at *TERT* and *SDHD* in melanoma respectively recruit or abrogate GABP transcription factor binding	37
Chapter 3	A common intronic variant of PARP1 confers melanoma risk via regulation by RECQL of an allelic DNA structure	85
Chapter 4	Global profiling of protein-DNA and protein-nucleosome binding affinities using quantitative mass spectrometry	147
Chapter 5	Discussion	191
Chapter 6	Summary Samenvatting	201 203
Chapter 7	Acknowledgments	207
Chapter 8	Curriculum vitae	213

Chapter 1

Introduction

Well then, I think the answer is that a circle has no beginning.
-*Harry Potter and the Deathly Hallows*, J.K. Rowling

A general introduction

The major epochs in the history of biology have led towards a preference for quantitative principles founded on a physical understanding of biological systems. Biology, as it was practiced in the 18th century, was defined in large part by *systematics*, the classification of various kinds of biological entity into their respective groups with respect towards their phenotypic differences. But rather than emphasize distinctions, biology of the 19th century asserted in contrast that *all* life is fundamentally *physicochemical*, in conjunction with developments in chemistry and the physical sciences at the time. Now, the dominant paradigm of biology in the 20th and 21st centuries has been the *informational* basis of life, that biological phenotype is digitally encoded in an organism's DNA molecules. Watson and Crick's foundational insight, the structure of this genetic information encoded in molecular form (DNA), is perhaps the single most disseminated achievement in molecular biology[1-3]. Yet since then, biologists have been laboring to unravel the mechanisms by which one "genome" is sufficient to produce essentially every component of a functional cell and, moreover, how these components are utilized differentially to specify the multiplicity of individual cell types that constitute a complex, multi-cellular organism. Indeed, the discovery by Jacob, Monod and others that expression of gene transcripts and thereby proteins is *regulated* and this regulation is sensitive to environmental changes represents another pillar of contemporary molecular biology[4-6]. The realization that gene expression, unlike the underlying genetic information itself, is defined contextually solved the problem of the specification of diverse cellular phenotypes from a single genomic state. An unspecified cell could "differentiate" into a more defined state on the basis of its internal state (i.e., the combination of all differentially expressed biomolecules present in that cell at a certain time) and the respective changes in the abundances of those biomolecules dictated by external stimuli. However, this immediately begs the question: how are these changes in *regulation*, in transcription, enacted? If the level of expression of a gene is subject to change, what factors mediate that change? The current biological consensus, and the overarching topic of this thesis, is that gene expression is predominantly regulated by cis- or trans-acting DNA-binding *transcription factors* and by *chromatin* state, the three-dimensional *epigenetic* structure of

the underlying DNA, which is itself regulated by DNA-binding chromatin remodeling and modifying enzymes.

Though the defining event of 21st century biology is almost indisputably the decoding of the human genome, following closely behind is the more general movement towards a *systems biology* perspective. This movement has been spurred by the sheer complexity uncovered in such transcriptional regulatory mechanisms over the last fifty-odd years. In general, the mechanistic basis of biological information translation is increasingly embedded within a holistic network framework that encompasses the entirety of the cell's or organism's components. New paradigms often stem from new technologies, and the field of biology is no exception to this principle. Advances most notably in next generation sequencing (NGS) technologies and workflows and also, most relevantly to this thesis, proteomics and particularly mass spectrometry-based proteomics have contributed heavily to a new *big data*, *multi-omics* biology[7]. Although the intrinsic meaningfulness of these terms is occasionally overstated, the basic premise is quite simple: to assess a complex systems with many interacting components, many measurements must be performed, and each measurement should evaluate many components. Although this approach is often fundamentally descriptive (or, in other words, *hypothesis generating*), useful biological inferences can often be made on the basis of genome-wide, proteome-wide, etc. datasets. For example, ChIP-seq experiments[8,9] simply describe the genomic locations of histone marks or the binding sites of DNA binding proteins genome-wide, yet from these genomic locations, potential gene regulatory relationships can be inferred. Interaction proteomics[10] experiments describe all the binding partners of a given bait (protein, DNA, small molecule, etc.) proteome-wide, yet from these binding partners, potential higher-order multimeric complexes and co-regulatory relationships can be inferred. A continually open question in scientific research is how novel and existing technologies can be best leveraged to solve outstanding problems. This thesis will address the technological question of how –omics approaches, specifically proteomics approaches, can shed new insights into problems in transcriptional regulation, cancer genomics, chromatin biology, and protein biochemistry.

In the next sections, I will provide a more detailed introduction to the biological and technological components of this thesis. These include but are not necessarily limited to:

1) Transcriptional regulation by transcription factors
2) The relationship between transcription and chromatin state
3) The regulatory potential of DNA sequence, structure, and motif architecture
4) Cancer genomics, cancer-associated DNA variants, and deregulation of gene expression in cancer
5) Mass spectrometry for studying protein-DNA and protein-nucleosome interactions

Transcriptional regulation by transcription factors

While Pardee, Jacob, and Monod[5,6] first identified a protein component that functionally regulated the expression of a different protein component, it has since been realized that factors with such capacity, transcription factors (TFs), are a large (~1600 proteins in humans) and diverse class of proteins[11]. In their simplest conception, TFs are proteins that simply influence the expression of gene products i.e. other proteins. However, in general, the current conception of TFs refers to *sequence-specific* DNA binding factors that regulate *transcriptional* processes. TF regulation may be direct through recruitment of RNA polymerases[12] or basal transcription machinery[13], or indirect through recruitment of chromatin remodeling or modifying enzymes[14]. Similarly, TFs may regulate transcription via sequence-specific genomic binding that is nearby to the regulated gene or quite far away [15]. The large number of TFs are sub-categorized by the presence of characteristic *DNA binding domains* (DBDs). There are often many homologous members within a *TF family* defined by a particular ancestral DBD. These homologous family members regularly have similar sequence preferences defined by their DBDs, yet individual factors often vary slightly on the theme of the consensus family *motif*. Different DBD families can have characteristic multimerization preferences, and this combinatorial action further expands the library of TF regulatory potential. For example, C2H2 zinc fingers (ZF) often bind as monomers, basic helix-loop-helix (bHLH) or basic leucine zipper (bZIP) are obligate homo- or hetero-dimers, and even higher order multimers are observed in some TF families[11]. Alternatively, TF families can utilize both homomeric and heteromeric binding modes. Within the human ETS family, for instance, nearly all members recognize a similar sequence motif and have the ability to do so monomerically, yet the heterodimer GABP (alpha and beta subunits) recognizes a similar motif

as an obligate multimer[16-18]. As an example of these various complementary modes of TF-dependent transcriptional regulation of a target gene, I will briefly introduce a case of *sequence-specific* transcriptional regulation of human telomerase (*TERT*) expression by a TF.

Numerous sequence variants, both inherited variants (germline) in the *TERT* locus[19-21] and acquired (somatic) variants in the *TERT* promoter [22-25], have previously been associated specifically with melanoma and many other cancers. In short, a causal variant (rs36115365) for the germline multi-cancer *TERT* risk locus was associated with differential, sequence variant-specific binding of the C2H2-ZF TF ZNF148[26]. Although, relevantly for this thesis, the identity of ZNF148 as a sequence-specific TF was initially established by mass spectrometry based workflows, a bevy of biochemical assays and cell-based assays were necessary to confirm that ZNF148 indeed: a) directly bound rs36115365 DNA b) bound to rs36115365 *in vivo* c) increased *TERT* expression allele-specifically d) had a downstream effect on telomere length via regulation of *TERT* expression. Intriguingly, the rs36115365 is located approximately 18kb upstream of the 5' end of the *TERT* gene and is, in fact, more proximal to the CLPTM1L gene. Therefore, ZNF148 regulation of *TERT* expression is an example of long-range transcriptional regulation where, importantly, assigning a regulatory interaction to the nearest gene in 1D sequence space would implicate the incorrect gene-variant relationship. As such, ZNF148 acts at a transcriptional enhancer of *TERT* expression and lies at a genomic location epigenetically marked by the characteristic H3K27ac/H3K4me1 histone post-translational modifications (PTMs). Chapter 2 of this thesis describes a related example of somatic variants in the *TERT* promoter inducing sequence-specific transcriptional up-regulation by the TF GABP[27]. The biological and oncological relevance of transcriptional *TERT* reactivation will be discussed in greater detail later in this introduction.

The relationship between transcription and chromatin state

The connection between transcriptional regulation by TFs at an *enhancer* or *promoter* element and the presence of defined histone PTMs or DNA modifications warrants a more detailed treatment, particularly because it is deeply connected to the idea of *chromatin state* as a major component of gene expression regulation[28-32].

The need for a model of DNA packing and compaction results directly from two simple observations: that DNA molecules carry a large negative charge, and that the molecular size of a "genome" is typically large relative to a cell's or a nuclei's internal volume. Electrostatic repulsion between DNA molecules and strands obviates the need for an active packing mechanism. In fact, Roger Kornberg proposed in 1974 that chromatin structure consisted of discrete, repetitive protein-DNA units, a DNA "super-coil" wrapped around an octamer of positively charge histone proteins[33]. In Kornberg's model, these "repeating units" could then flexibly assemble into a higher-order chromatin *fiber*. A landmark study in 1984 by Tim Richmond, Daniela Rhodes, Aaron Klug, and colleagues solved the structure of this super-helical DNA-histone complex, termed the *nucleosome*, at intermediate 7Å resolution[34]. This result was improved to an atomic resolution of 2.8Å by Karolin Luger working with Tim Richmond in 1997[35]. Further structural and functional studies focused on the higher-order assembly of individual nucleosomes into the so-called 30nM chromatin fiber. Reported 30nM fiber structures have been somewhat discrepant. Tim Richmond and colleagues solved the intermediate 9Å structure of a tetra-nucleosome[36], which, in combination with additional functional and electron microscopy data, led them to propose a two-start helical model for the 30nM fiber[37]. In contrast, Daniela Rhodes and colleagues proposed an interdigitated one-start helical model, highlighting the importance of linker histone H1 and longer linker lengths between nucleosome units[38,39]. These different structural models, possibly, reflect independent and functional chromatin states[40]. However, a beautiful and more recent study suggested that a two-start double helical 30nM chromatin fiber forms in the presence of linker histone H1 and low salt, and is not strongly dependent on DNA linker length[41]. Perhaps more interestingly, whether or not the 30nM chromatin fiber actually forms *in vivo*, and if so in what conditions, is a topic of ongoing discussion[42,43]. A recent study, utilizing fluorescent dye enabled deposition of contrast enhancing polymers for electron microscopy tomography, concluded based on novel *in vivo* chromatin imaging data that the 30nM fiber is not observed in interphase or mitotic chromatin[44]. Instead, the authors claim, human cells organize their chromatin as a 5-24nM flexible bead-like polymer, where only nucleosome *concentration* dictates chromatin compaction, as opposed to any structured, higher-order chromatin fiber folding.

Though more work is necessary to comprehensively resolve the details of chromatin folding *in vivo*, the relevance of chromatin compaction to gene expression regulation is clear: chromatin structure can represent a substantial steric block for the transcriptional machinery[45]. Indeed, the earliest description of distinct *chromatin states*, by Heitz in 1928, distinguished two main chromatin states based on their level of compaction[46]. Chromatin compaction, it has since emerged, is intrinsically related to transcriptional potential in the underlying genetic region; *euchromatin*, decondensed chromatin, is gene-rich, accessible, and transcriptionally active, while *heterochromatin*, tightly condensed chromatin, is gene-poor, inaccessible, and transcriptionally silent[47]. In fact, the relationship between chromatin condensation state and transcription is so strong that heterochromatin must be actively *remodeled*, oftentimes an ATP-consuming process, to facilitate transcription[48]. On the one hand, this remodeling is enacted by a sub-class of TFs, commonly referred to as *pioneer factors*[49-51]. Pioneer transcription factors are, by definition, TFs that have the capacity to recognize their DNA motif in the context of a nucleosome particle. Key pioneer TFs can partially recognize exposed "half-motifs" along the nucleosomal super-helix[52]. However, cell- and tissue-specificity in pioneer TF site remodeling can also depend on the presence of cooperative interactions with other TF co-factors[53,54]. On the other hand, nucleosome and chromatin remodeling by pioneer TFs is correlated with modifications to *histone post-translational modifications (PTMs)* at those loci[48,55]. Indeed, eu- and heterochromatin can be distinguished by characteristic histone tail PTMs (i.e., H3K27 acetylation at euchromatin v. H3K9 tri-methylation at heterochromatin). With few exceptions[56], histone PTMs do not affect the core structure of the nucleosome[57]. Rather, histone modifications seem to affect nucleosome stability, dynamics, higher-order chromatin folding, and interactions with histone chaperones or chromatin remodellers[58,59]. Foundational papers for the field of *epigenetics* established the regulatory and phenotypic importance of histone tails (and thereby histone tail PTMs)[60] and the enzymes that deposit and remove them[61]. The paradigm of histone tail PTM "readers, writers, and erasers" has recently culminated in an intense effort to identify, proteome-wide, functional PTM-specific proteins and protein complexes[62-65]. These chromatin modifiers and remodelers, through steric regulation via nucleosome positioning and chromatin state regulation via histone PTM modifications, subsequently regulate transcription in development

and cell differentiation[66]. Thus, the initiation and maintenance of chromatin states is a complex and multi-faceted process that facilitates specification of diverse cell states via differential gene expression regulation.

The regulatory potential of DNA sequence, structure, and motif architecture

Although I have, so far, considered the regulatory potential of DNA itself purely in the context of its 1D sequence (and thereby its interactions with sequence-specific TFs), DNA is of course a molecule that exists in 3D space. Correspondingly, recent research has revealed surprising structural aspects of DNA elements that influence their gene regulatory function. First, like histone proteins, DNA can be post-translationally modified, predominantly via CpG methylation, with concurrent effects on gene expression regulation[67]. DNA methylation was discovered by chemical means as early as 1948[68], though its generally repressive effect on gene expression was discovered relatively later[69,70]. In line with its gene regulatory function, CpGs are depleted genome-wide in mammalian genomes[71], yet unmethylated CpG regions preferentially associate with genes[72]. Similarly, mCpGs are predominantly located in heterochromatic regions, and they co-localize with and are mechanistically linked to the heterochromatic H3K9me3 histone mark[73]. Aberrant gene-specific CpG (hypo)-methylation has been associated with cancer in human tumors[74]. The identification of the regulatory function of mCpG has, as with histone PTMs, led to the identification of mCpG specific "reader" proteins including MeCP2[75]. Proteome-wide approaches have greatly expanded the known repertoire of mCpG (and its chemical derivatives) specific reader proteins[76].

In addition to functional modulation by epigenetic modifications, the DNA molecule itself has a shape component, which is related to but also distinct from its sequence component. Ground-breaking work from Remo Rohs and colleagues demonstrated that shape features of dsDNA, including minor groove width, roll, propeller twist, and helix turn[77], can improve predictive models of TF DNA binding[78-81]. Such shape-based features have recently been applied to epigenetic data, suggesting that DNA methylation affects local DNA shape and thereby protein binding[82] (although structural data shows that local shape changes are modest and global DNA shape changes are absent[83,84]). Binding

models considering histone modifications may improve TF DNA binding predictions for some families of TF[85]. Furthermore, not only DNA shape but also allosteric effects induced by binding at distal DNA elements can affect protein-DNA binding[86,87]. Moreover, while Watson and Crick's famous solution of the DNA structure[1] represents the *B-form* of DNA, double-stranded DNA (dsDNA) can adopt alternate conformations including A-DNA and Z-DNA[88]. Beyond dsDNA forms, DNA can adopt a variety of more exotic single-stranded (ssDNA), triplex (three-stranded), or quadruplex (four-stranded) structures[89]. Although some of these structures, most notably *G-quadruplexes (G4s)*, have been associated with various modes of gene expression regulatory function, the role of DNA structures in mediating gene expression, if any, is still an active area of research[90-93]. Chemical approaches stabilizing G4 structures with small molecule ligands have been associated with transcriptional changes in nearby genes, for example, downregulation of *c-Myc* expression upon G4-stabilizing ligand treatment[93,94]. Concomitantly, G4-stabilizing molecules induce local epigenetic reprogramming including the removal of H3K4me3 (a mark of active promoters) and induction of heterochromatic H3K9me3 and DNA CpG methylation[95]. More recently, NGS approaches have facilitated the mapping of G-quadruplexes, non-B DNA, and even RNA-DNA triplexes (R-loops) genome-wide, leading to the observation that many non-canonical DNA structures map to regulatory regions (as distinguished by epigenetic marks)[96-100]. Because many techniques and reagents for studying DNA and RNA secondary structures have emerged relatively recently, however, protein mediators of any regulatory function assigned to such secondary structures remain largely uncovered. Many reported DNA and RNA G4 binding proteins are helicases related to G-quadruplex unwinding[101,102]. Although there are a few reports of G-quadruplex interactions with chromatin remodeling[103,104] or modifying enzymes[105-107], a thorough explanation for the connection between G-quadruplex structure and chromatin state is lacking. Chapter 4 of this thesis presents data suggesting that stable G-quadruplexes might act as recognition motifs for some chromatin remodeling and modifying enzymes. Similarly, although it seems clear that G-quadruplexes and non-canonical DNA structures have some link to cancer development and progression[108,109], possibly via effects on genome maintenance and instability related to DNA repair and chromatin modulation[110-113], this area of research is relatively immature. Chapter 3 of this

thesis presents an interesting case of allele-specific DNA secondary structure formation in a cancer-associated insertion/deletion variant correlated with differential helicase binding and gene expression at the target locus[114].

Cancer-associated DNA variants, cancer genomics, and deregulation of gene expression in cancer

It is clear that there exists great complexity among the mechanisms by which a cell might differentially regulate the expression of genes. This complexity is, on the one hand, a benefit, because it facilitates an incredible diversity of cell phenotypes, cell responses to perturbation, and underlying gene regulatory mechanisms and networks. On the other hand, cancer development proceeds via the subversion of normal, healthy cellular growth pathways and the gene regulatory networks that maintain them to produce an unrestrained, immortalized growth phenotype. In a sense, cancer is an evolutionary disease[115]. Selection at the level of the (eukaryotic, multi-cellular) individual selects for cancer suppressive or resistant genotypes, in order to protect the life and reproductive potential of the organism as a whole. Selection at the level of the cell selects for, simply, propagation. Therefore, complex gene regulatory mechanisms have the disadvantage of presenting numerous distinct opportunities for the cell to acquire advantageous growth phenotypes. It has been proposed that the cancer phenotype can be distinguished by the acquisition of a few key molecular and cellular traits, famously referred to as the "hallmarks of cancer"[116,117]. In many ways, the "hallmarks of cancer" paradigms has shaped and defined cancer research for the last fifteen odd years, and progress has been made towards elucidating the mechanisms behind, and proposing targeted therapeutic interventions for, each of the cancer hallmarks. For example, *BRAF* mutations (V600E) are frequent in a variety of cancers, particularly melanoma and are associated with constitutively active growth signaling via the MAP kinase pathway[118]. And indeed, targeted therapeutic options for *BRAF* V600E (i.e., vemurafenib) exist and are in clinical use[119]. However, in line with the complexity and resiliency of oncogenic pathways, acquired resistance to vemurafenib treatment is common. Furthermore, paradoxically, targeted BRAF inhibition is often associated with MAP kinase signaling *activation* and, as such, can induce additional cutaneous lesions. Similarly, it has become

clear that *BRAF* mutations are not the only molecular mechanism by which cancer cells can constitutively activate MAP kinase signaling pathways[120,121]. In sum, despite the considerable progress that has been made in understanding the mechanisms driving cancer formation, more research is still necessary to outsmart cancer via novel clinical avenues.

While coding mutations in "driver" oncogenes or "tumor suppressor" genes have received much study over the past two decades[122,123], especially with the advent of NGS technologies and cancer genomics studies, the role of *non-coding* mutations in cancer has been of more recent interest[124]. Large-scale cancer genomics sequencing studies have identified genome-wide mutational signatures that, in some cases, were cancer specific[125]. Additional studies implicated "hotspots" for non-coding somatic mutations, most often in promoter regions or 5' untranslated regions (UTRs)[126,127]. Of interest, hotspot somatic mutations in the *TERT* promoter were the most consistently and significantly mutated non-coding region across all cancers, in line with previous findings[22-24]. Certainly, it should not be surprising that promoter mutations associated with *TERT* reactivation are the most frequent significant of all non-coding variation. Indeed, *TERT* reactivation represents a classic "hallmark of cancer" by enabling replicative immortality. Without *TERT* expression or some other telomere elongating mechanism, gradual telomere shortening and eventual senescence are an insurmountable obstacle to oncogenesis[128]. Therefore, cancer cells are under intense selective pressure to activate telomerase expression, or activate an Alternative Lengthening of Telomeres pathway[129]. Because *TERT* expression is generally silenced in differentiated cells, activating coding mutations in the telomerase protein are not often seen. Mechanisms of *TERT* reactivation are, therefore, often by non-coding means such as promoter mutations or gene amplifications (indeed, it is currently thought that most cancer-associated non-coding variants act via sequence-specific TF binding and allele-specific transcriptional regulatory mechanisms). A pan-cancer cancer genomics analysis highlighted the significance of mutations in the *TERT* promoter[130]. It was noted very early on that both of the most frequent *TERT* promoter mutations create non-endogenous ETS TF family binding motifs. Indeed, as demonstrated in Chapter 2 of this thesis, the ETS factor GABP binds heterotetramerically to the *TERT* promoter via both endogenous ETS motifs and non-endogenous, mutation-specific ETS motifs simultaneously, thereby activating mono-allelic

TERT expression and inducing an epigenetic switch marked by active histone PTMs[27,131,132]. GABP activation of *TERT* is additionally related to BRAF V600E signaling via aberrant FOS activation[133]. *TERT* reactivation via ETS family motifs led researchers to seek mutation-specific changes to ETS family motifs in non-coding hotspot mutations near other genes. It was suggested that promoter mutations at the *SDHD* gene might also downregulate *SDHD* expression via the loss of ETS factor binding, and in this case the ETS factor ELF1 was proposed as the likely causal factor based on bioinformatics analysis[127]. However, later research, also shown in Chapter 2 of this thesis, demonstrated that mutation-specific loss of GABP binding, as with *TERT* promoter mutations, was largely responsible for the reduction of *SDHD* expression in mutated tumors[134]. Finally, in addition to coding driver mutations and somatic non-coding variants, the explosion of *genome-wide association studies (GWAS)* over the last decade has identified an overwhelming number of germline variants that have been associated with hereditary risk for many diseases, including cancers[135,136]. In melanoma alone, there are ~20 loci associated with hereditary risk[20]. Each locus often contains tens or hundreds of potential causal variants. Therefore, unraveling the molecular mechanism by which a locus confers risk for the disease is so far often performed on a case-by-case basis and is highly non-trivial[114]. Ultimately, the functional annotation of non-coding variants in cancer, both germline and somatic, will be a major direction for concerted future effort.

Mass spectrometry for studying protein-DNA and protein-nucleosome interactions

For profiling the function of non-coding, cancer-associated sequence variants, there exist many classic molecular and cellular biology experimental techniques. Among these, techniques such as luciferase assays and Pol II chromatin immunoprecipitation followed by qPCR (ChIP) can monitor variant-specific transcriptional activity; EMSA, affinity purification followed by western blot, or ChIP-qPCR can identify variant specific protein binding; isothermal titration calorimetry (ITC), fluorescence polarization (FP) or Förster resonance energy transfer (FRET), surface plasmon resonance (SPR), and EMSA are biochemical assays that can directly assess the allele-specific affinity of protein-DNA interactions. High-throughput approaches for determining

protein-DNA interaction specificities have attracted recent attention[137]. Although these techniques are powerful and have been widely utilized in many functional genomics studies, their general disadvantage is that they are *targeted*. In contrast, mass spectrometry based interaction proteomics has emerged as an *unbiased* tool for identifying specific protein interactions, including variant-specific binding of proteins to DNA[10,138,139]. Mass spectrometry based workflows have yielded large-scale protein-protein interaction networks at the level of entire proteomes, or thousands of individual baits[140-142]. Cross-linking mass spectrometry (XL-MS) workflows, similarly, are now capable of measuring thousands of residue-level interactions *in vivo*[143,144]. However, methods and applications of mass spectrometry for studying protein-DNA interactions have received less attention, despite the potential of mass spectrometry for providing insights into such regulatory interactions. Butter, Mann, and colleagues provided an initial demonstration that quantitative mass spectrometry could be used for the identification of DNA-binding proteins and, perhaps more critically, the identification of *sequence-specific* protein-DNA interactions[145,146]. Additional studies identified specific interactors of ultra-conserved DNA elements[147] and interaction-based evolution of the telomere-protecting shelterin complex[148]. Protein-RNA interactions have been assayed using mass spectrometry using similar workflows, or UV-crosslinking based workflows[149-151]. Spruijt et al. identified modification and differentiation stage-specific interactors of mCpG DNA and its various chemical derivatives including hydroxyl-methylation[76]. Edupuganti et al. performed a conceptually similar studying, identifying and functionally profiling specific readers of N6-methyladenosine (m6A) modified RNA[152]. In addition to epigenetic DNA modifications, specific readers of histone PTMs in a histone tail peptide context[62,64] and a nucleosome context[63] have been identified by mass spectrometry. Fang et al. used a mass spectrometry approach to identify variant specific TFs for a common, multicancer-associated SNP in the *TERT* locus[26]. In Chapter 2 of this thesis, Makowski et al.[27] and Zhang et al.[134] used mass spectrometry to identify somatic variant specific TFs, observing gain of GABP binding and upregulation of *TERT* or loss of GABP binding and downregulation of *SDHD*, respectively. Chapter 3 of this thesis uses mass spectrometry with an insertion/deletion DNA bait to detect specific protein binding associated with transcriptional regulation of the *PARP1* gene in melanoma[114]. Though these relatively few examples demonstrate the promise of

mass spectrometry applied to protein-DNA interactions, most importantly for uncovering possible regulatory interactions, the massive number of putatively functional disease-associated variants demands much additional work.

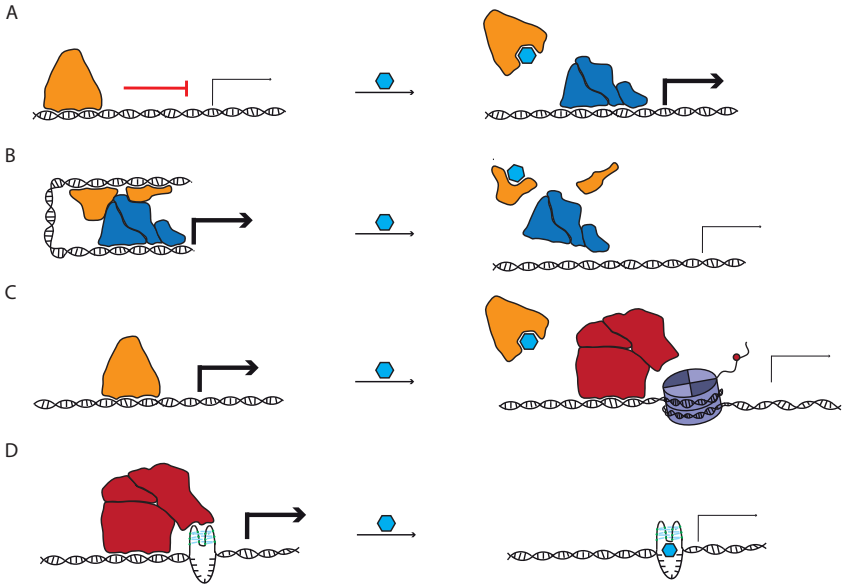


Figure 1. Modes of transcriptional gene expression regulation

A Canonical gene expression regulation. A single repressor transcription factor (TF) regulates the expression of a target gene. In the presence of a stimulatory ligand, the repressor is sterically blocked from DNA binding and thus transcription can commence.

B Long-range gene regulation. Regulatory TFs can bind in trans and activate transcription of a target gene from long distances (tens or hundreds of kilobases). If the stabilizing effect of this long-distance interaction is lost due to inhibition of regulatory TFs, the basal transcriptional machinery binds inefficiently and gene expression is decreased.

C Gene expression regulation by chromatin remodellers and modifiers. If activating TFs are inhibited from binding, chromatin remodelling or modiffying enzymes might bind instead and remodel local chromatin state, thereby using nucleosomes to compact DNA and sterically repress expression of the target gene.

D Gene expression regulation by DNA structural elements. DNA secondary structural elements may recruit regulatory TF complexes, which activate target gene expression. Inhibition of this DNA structure-TF complex interaction by small molecule ligands subsequently decreases expression of the target gene.

[Note: In this figure, I generally refer to “activating” and “repressive” modes of gene expression regulation in mutually exclusive terms. However, it is well established and should be noted that almost all modes of gene expression regulation can both activate and repress gene expression depending on the context.]

Outline of this thesis

This thesis argues, in the form of a few case studies, that the application of mass spectrometry to the study of protein-DNA interactions represents a powerful and versatile tool for uncovering potential mechanisms of transcriptional gene expression regulation, particularly aberrant gene regulation in the context of cancer.

Chapter 2 shows examples of somatic (promoter) mutation specific binding by the GABP protein. Using mass spectrometry, we see that recurrent promoter mutations in the *TERT* promoter are associated with novel ETS family TF motifs that subsequently induce GABP binding and thereby transcriptional *TERT* reactivation. However, intriguingly, recurrent promoter mutations at the *SDHD* gene *disrupt* native ETS motifs and thus decrease *SDHD* expression. Based on motif architecture and AP-MS experiments, our data suggests that activating GABP binding at *TERT* uses a tetrameric mode, while wild-type binding at *SDHD* utilizes a dimeric mode. Therefore, despite the identification of a single factor (GABP) associated with mutation specific binding at two recurrently mutated promoters, the binding mechanism leading to oncogenic transcriptional deregulation is entirely different.

My contributions to this chapter included conceiving of and designing the TERT study with M.V. and K.M.B., performing DNA binding assays including EMSA and western blot experiments with E.W., performing, measuring, and analyzing mass spectrometry experiments, writing the *TERT* manuscript with input from all authors, and contributing to the writing of the SDHD manuscript by T.Z. and K.M.B.

In **Chapter 3**, we identify allele-specific regulatory potential and allele-specific protein binding at a common, germline variant in the melanoma-associated *PARP1* risk locus. The variant is an insertion/deletion that falls inside a short hexameric repeat and therefore creates no novel sequence motifs. Instead, we observed allele-specific binding of a number of proteins annotated as recognizing DNA structural features including the DNA helicase RECQL, which regulates *PARP1* expression. Finally, we provide additional analysis of the structure-specific protein binding properties of this insertion/deletion variant using chemical structural perturbations.

My contributions to this chapter included molecular biology assays of candidate SNPs including EMSA and luciferase assays, DNA binding assays

and dimethyl labeling and TMT mass spectrometry experiments, analyzing mass spectrometry data and making figures, and contributing to the writing of the manuscript by J.C. and K.M.B.

Mass spectrometry workflows described in Chapter 2 & 3 use relative, semi-quantitative labeling to identify sequence-specific protein binding. **Chapter 4** describes an absolutely quantitative mass spectrometry workflow for assessing protein-DNA or protein-nucleosome apparent binding affinities (K_d^{App}). Titrated DNA sequences or nucleosomes are used for affinity purifications, after which isobaric TMT labeling and multiplexed mass spectrometry analysis are performed. After performing benchmarking studies with the consensus SP/KLF motif, we perform a larger survey of a canonical set of ssDNA and dsDNA motifs. Notably, we observe a number of chromatin remodeling and modifying complexes, including SWI/SNF, ISWI, PRC2, and NuRD, binding with high affinity to a reported DNA G-quadruplex forming sequence from the *c-Myc* promoter. These interactions appear to be somewhat structurally specific, as chemical perturbation of the G-quadruplex structure resulted in a decreased DNA affinity for many of these complexes. Further, we observe high affinity interactions between SWI/SNF and ISWI subunits and modified and unmodified nucleosomes. Intriguingly, we see high affinity binding for catalytic SWI/SNF subunits even in the absence of H3 modifications; however, accessory SWI/SNF subunits bind to the nucleosome with high affinity only in the presence of H3K9AcK14Ac. We foresee fully quantitative binding analysis providing a link between absolute, copy-number proteome analysis and transcriptional output as measured by absolutely quantitative RNA-seq workflows.

My contributions to this chapter included conceiving of and designing the study with M.V., performing DNA binding assays including EMSA and western blot experiments with C.G., performing DNA cross-linking and recombinant protein experiments, performing, measuring, and analyzing mass spectrometry experiments with C.G., analyzing other data and making figures, and writing the manuscript with input from all authors.

In the **Conclusion**, I discuss more informally the main research findings of this thesis and offer a broad perspective on future research into protein-DNA interactions using mass spectrometry.

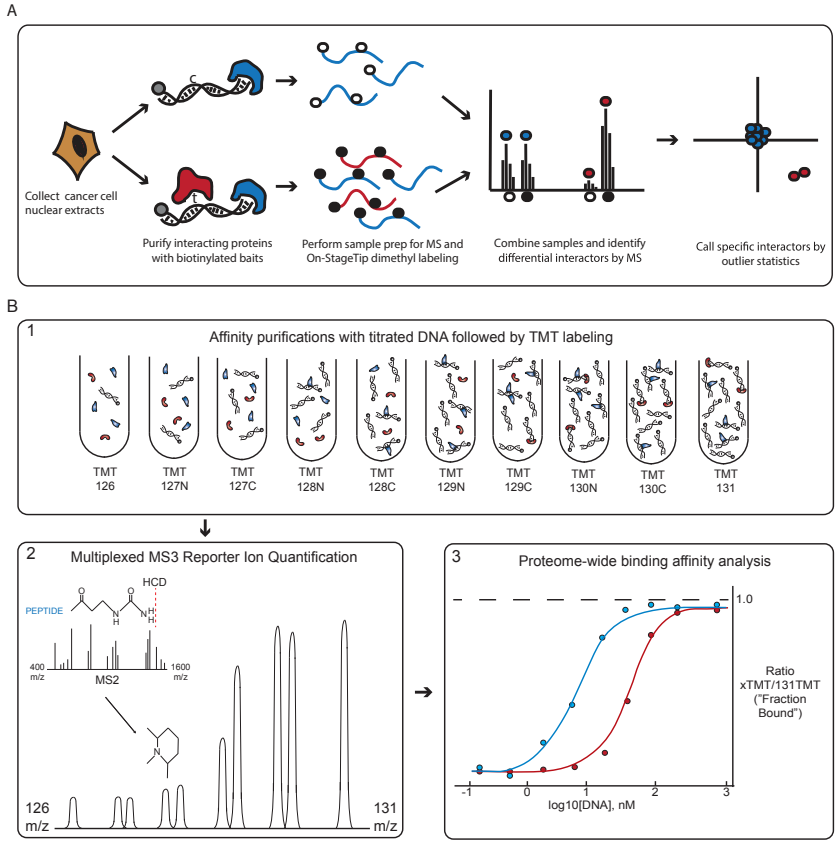


Figure 2 Mass spectrometry based methods for studying protein-DNA and protein-nucleosome interactions

A Semi-quantitative analysis of sequence-specific protein-DNA interactions. Nuclear proteins are isolated from cancer cell lines and are enriched by affinity purification using biotinylated DNA oligonucleotides encompassing either a wild-type or a variant sequence. Proteins are digested to peptides, and these peptides are chemically labelled, most often using dimethyl chemical labeling. Binding ratios are calculated based on relative quantification in the MS1 spectra of the isotopically labelled peptides. Outliers are called from a background cloud of non-specifically binding proteins.

B Absolute quantification of protein-DNA or protein-nucleosome KdApp values. A series of DNA oligonucleotide or nucleosome affinity purifications is performed as above, using a titrated bait of known concentration. Bound proteins are digested to peptides and labelled with isobaric TMT labels. Mass spectrometry analysis of the multiplexed peptide mixture is performed. Protein ratios are quantified from peptide reporter ions analyzed in MS3 spectra. KdApp values are quantified by fitting measured protein ratios using a Hill-like curve.

References

1 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

2 Wilkins, M. H., Stokes, A. R. & Wilson, H. R. Molecular structure of deoxypentose nucleic acids. *Nature* **171**, 738-740 (1953).

3 Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740-741 (1953).

4 1965 nobel prize for medicine and physiology. *Triangle* **7**, 164 (1965).

5 Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356 (1961).

6 Pardee, A. B., Jacob, F. & Monod, J. The genetic control and cytoplasmic expression of "Inducibility" in the synthesis of β-galactosidase by E. coli. *Journal of Molecular Biology* **1**, 165-178, doi:https://doi.org/10.1016/S0022-2836(59)80045-0 (1959).

7 Joyce, A. R. & Palsson, B. O. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **7**, 198-210, doi:10.1038/nrm1857 (2006).

8 Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572-576, doi:10.1038/nature05632 (2007).

9 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837, doi:10.1016/j.cell.2007.05.009 (2007).

10 Smits, A. H. & Vermeulen, M. Characterizing Protein-Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends Biotechnol* **34**, 825-834, doi:10.1016/j.tibtech.2016.02.014 (2016).

11 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665, doi:10.1016/j.cell.2018.01.029 (2018).

12 Frietze, S. & Farnham, P. J. Transcription factor effector domains. *Subcell Biochem* **52**, 261-277, doi:10.1007/978-90-481-9069-0_12 (2011).

13 Koster, M. J., Snel, B. & Timmers, H. T. Genesis of chromatin and transcription dynamics in the origin of species. *Cell* **161**, 724-736, doi:10.1016/j.cell.2015.04.033 (2015).

14 Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707-719, doi:10.1016/j.cell.2007.01.015 (2007).

15 Spitz, F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin Cell Dev Biol* **57**, 57-67, doi:10.1016/j.semcdb.2016.06.017 (2016).

16 LaMarco, K., Thompson, C. C., Byers, B. P., Walton, E. M. & McKnight, S. L. Identification of Ets- and notch-related subunits in GA binding protein. *Science* **253**, 789-792 (1991).

17 Thompson, C. C., Brown, T. A. & McKnight, S. L. Convergence of Ets- and notch-related structural motifs in a heteromeric DNA binding complex. *Science* **253**, 762-768 (1991).

18 Rosmarin, A. G., Resendes, K. K., Yang, Z., McMillan, J. N. & Fleming, S. L. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. *Blood Cells Mol Dis* **32**, 143-154 (2004).

19 Law, M. H. *et al.* Meta-analysis combining new and existing data sets confirms that the TERT-CLPTM1L locus influences melanoma risk. *J Invest Dermatol* **132**, 485-487, doi:10.1038/jid.2011.322 (2012).

20 Law, M. H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat Genet* **47**, 987-995, doi:10.1038/ng.3373 (2015).

21 Rafnar, T. *et al.* Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet* **41**, 221-227, doi:10.1038/ng.296 (2009).

22 Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-961, doi:10.1126/science.1230062 (2013).

23 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-959, doi:10.1126/science.1229259 (2013).

24 Killela, P. J. *et al.* TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc Natl Acad Sci U S A* **110**, 6021-6026, doi:10.1073/pnas.1303607110 (2013).

25 Heidenreich, B., Rachakonda, P. S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Curr Opin Genet Dev* **24**, 30-37, doi:10.1016/j.gde.2013.11.005 (2014).

26 Fang, J. *et al.* Functional characterization of a multi-cancer risk locus on chr5p15.33 reveals regulation of TERT by ZNF148. *Nat Commun* **8**, 15034, doi:10.1038/ncomms15034 (2017).

27 Makowski, M. M. *et al.* An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations. *Proteomics* **16**, 417-426, doi:10.1002/pmic.201500327 (2016).

28 Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**, 245-254, doi:10.1038/ng1089 (2003).

29 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705, doi:10.1016/j.cell.2007.02.005 (2007).

30 Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412, doi:10.1038/nature05915 (2007).

31 Rando, O. J. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr Opin Genet Dev* **22**, 148-155, doi:10.1016/j.gde.2012.02.013 (2012).

32 Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol* **16**, 178-189, doi:10.1038/nrm3941 (2015).

33 Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-871 (1974).

34 Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 A resolution. *Nature* **311**, 532-537 (1984).

35 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).

36 Schalch, T., Duda, S., Sargent, D. F. & Richmond, T. J. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **436**, 138-141, doi:10.1038/nature03686 (2005).

37 Dorigo, B. *et al.* Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* **306**, 1571-1573, doi:10.1126/science.1103124 (2004).

38 Robinson, P. J., Fairall, L., Huynh, V. A. & Rhodes, D. EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proc Natl Acad Sci U S A* **103**, 6506-6511, doi:10.1073/pnas.0601212103 (2006).

39 Routh, A., Sandin, S. & Rhodes, D. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc Natl Acad Sci U S A* **105**, 8872-8877, doi:10.1073/pnas.0802336105 (2008).

40 Robinson, P. J. & Rhodes, D. Structure of the '30 nm' chromatin fibre: a key role for the linker histone. *Curr Opin Struct Biol* **16**, 336-343, doi:10.1016/j.sbi.2006.05.007 (2006).

41 Song, F. *et al.* Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science* **344**, 376-380, doi:10.1126/science.1251413 (2014).

42 Maeshima, K., Hihara, S. & Eltsov, M. Chromatin structure: does the 30-nm fibre exist in vivo? *Curr Opin Cell Biol* **22**, 291-297, doi:10.1016/j.ceb.2010.03.001 (2010).

43 Fussner, E., Ching, R. W. & Bazett-Jones, D. P. Living without 30nm chromatin fibers. *Trends Biochem Sci* **36**, 1-6, doi:10.1016/j.tibs.2010.09.002 (2011).

44 Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357**, doi:10.1126/science.aag0025 (2017).

45 Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nat Rev Genet* **8**, 35-46, doi:10.1038/nrg2008 (2007).

46 Heitz, E. Das heterochromatin der Moose. Jb. Wiss. Bot. 69: 728-818. *View Article PubMed/NCBI Google Scholar* (1928).

47 Huisinga, K. L., Brower-Toland, B. & Elgin, S. C. The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma* **115**, 110-122, doi:10.1007/s00412-006-0052-x (2006).

48 Li, G. & Reinberg, D. Chromatin higher-order structures and gene regulation. *Curr Opin Genet Dev* **21**, 175-186, doi:10.1016/j.gde.2011.01.022 (2011).

49 Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends Genet* **27**, 465-474, doi:10.1016/j.tig.2011.07.002 (2011).

50 Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**, 2227-2241, doi:10.1101/gad.176826.111 (2011).

51 Drouin, J. Minireview: pioneer transcription factors in cell fate specification. *Mol Endocrinol* **28**, 989-998, doi:10.1210/me.2014-1084 (2014).

52 Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555-568, doi:10.1016/j.cell.2015.03.017 (2015).

53 Mayran, A. *et al.* Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nat Genet* **50**, 259-269, doi:10.1038/s41588-017-0035-2 (2018).

54 Donaghey, J. *et al.* Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nat Genet* **50**, 250-258, doi:10.1038/s41588-017-0034-3 (2018).

55 Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074-1080, doi:10.1126/science.1063127 (2001).

56 North, J. A. *et al.* Histone H3 phosphorylation near the nucleosome dyad alters chromatin structure. *Nucleic Acids Res* **42**, 4922-4933, doi:10.1093/nar/gku150 (2014).

57 Lu, X. *et al.* The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure. *Nat Struct Mol Biol* **15**, 1122-1124, doi:10.1038/nsmb.1489 (2008).

58 Tessarz, P. & Kouzarides, T. Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol* **15**, 703-708, doi:10.1038/nrm3890 (2014).

59 Bowman, G. D. & Poirier, M. G. Post-translational modifications of histones that influence nucleosome dynamics. *Chem Rev* **115**, 2274-2295, doi:10.1021/cr500350x (2015).

60 Kayne, P. S. *et al.* Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. *Cell* **55**, 27-39 (1988).

61 Brownell, J. E. *et al.* Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* **84**, 843-851 (1996).

62 Vermeulen, M. *et al.* Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142**, 967-980, doi:10.1016/j.cell.2010.08.020 (2010).

63 Bartke, T. *et al.* Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143**, 470-484, doi:10.1016/j.cell.2010.10.012 (2010).

64 Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M. & Mann, M. A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. *Mol Cell* **49**, 368-378, doi:10.1016/j.molcel.2012.10.026 (2013).

65 Eberl, H. C., Mann, M. & Vermeulen, M. Quantitative proteomics for epigenetics. *Chembiochem* **12**, 224-234, doi:10.1002/cbic.201000429 (2011).

66 Chen, T. & Dent, S. Y. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat Rev Genet* **15**, 93-106, doi:10.1038/nrg3607 (2014).

67 Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209-213, doi:10.1038/321209a0 (1986).

68 Hotchkiss, R. D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* **175**, 315-332 (1948).

69 Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226-232 (1975).

70 Compere, S. J. & Palmiter, R. D. DNA methylation controls the inducibility of the mouse metallothionein-I gene lymphoid cells. *Cell* **25**, 233-240 (1981).

71 Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8**, 1499-1504 (1980).

72 Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91-99 (1985).

73 Rose, N. R. & Klose, R. J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta* **1839**, 1362-1372, doi:10.1016/j.bbagrm.2014.02.007 (2014).

74 Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89-92 (1983).

75 Lewis, J. D. *et al.* Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**, 905-914 (1992).

76 Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146-1159, doi:10.1016/j.cell.2013.02.004 (2013).

77 Zhou, T. *et al.* DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* **41**, W56-62, doi:10.1093/nar/gkt437 (2013).

78 Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248-1253, doi:10.1038/nature08473 (2009).

79 Abe, N. *et al.* Deconvolving the recognition of DNA shape from sequence. *Cell* **161**, 307-318, doi:10.1016/j.cell.2015.02.008 (2015).

80 Mathelier, A. *et al.* DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst* **3**, 278-286 e274, doi:10.1016/j.cels.2016.07.001 (2016).

81 Yang, L. *et al.* Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol* **13**, 910, doi:10.15252/msb.20167238 (2017).

82 Rao, S. *et al.* Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding. *Epigenetics Chromatin* **11**, 6, doi:10.1186/s13072-018-0174-4 (2018).

83 Renciuk, D., Blacque, O., Vorlickova, M. & Spingler, B. Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic Acids Res* **41**, 9891-9900, doi:10.1093/nar/gkt738 (2013).

84 Hardwick, J. S. *et al.* 5-Formylcytosine does not change the global structure of DNA. *Nat Struct Mol Biol* **24**, 544-552, doi:10.1038/nsmb.3411 (2017).

85 Xin, B. & Rohs, R. Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res*, doi:10.1101/gr.220079.116 (2018).

86 Kim, S. *et al.* Probing allostery through DNA. *Science* **339**, 816-819, doi:10.1126/science.1229223 (2013).

87 Noy, A., Maxwell, A. & Harris, S. A. Interference between Triplex and Protein Binding to Distal Sites on Supercoiled DNA. *Biophys J* **112**, 523-531, doi:10.1016/j.bpj.2016.12.034 (2017).

88 Travers, A. & Muskhelishvili, G. DNA structure and function. *FEBS J* **282**, 2279-2295, doi:10.1111/febs.13307 (2015).

89 Kaushik, M. *et al.* A bouquet of DNA structures: Emerging diversity. *Biochem Biophys Rep* **5**, 388-395, doi:10.1016/j.bbrep.2016.01.013 (2016).

90 Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-780, doi:10.1038/nrg3296 (2012).

91 Varizhuk, A. *et al.* The expanding repertoire of G4 DNA structures. *Biochimie* **135**, 54-62, doi:10.1016/j.biochi.2017.01.003 (2017).

92 Hansel-Hertsch, R., Di Antonio, M. & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* **18**, 279-284, doi:10.1038/nrm.2017.3 (2017).

93 Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**, 8627-8637, doi:10.1093/nar/gkv862 (2015).

94 Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* **99**, 11593-11598, doi:10.1073/pnas.182256799 (2002).

95 Guilbaud, G. *et al.* Local epigenetic reprogramming induced by G-quadruplex ligands. *Nat Chem* **9**, 1110-1117, doi:10.1038/nchem.2828 (2017).

96 Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**, 877-881, doi:10.1038/nbt.3295 (2015).

97 Hansel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**, 1267-1272, doi:10.1038/ng.3662 (2016).

98 Kouzine, F. *et al.* Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. *Cell Syst* **4**, 344-356 e347, doi:10.1016/j.cels.2017.01.013 (2017).

99 Chen, L. *et al.* R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol Cell* **68**, 745-757 e745, doi:10.1016/j.molcel.2017.10.008 (2017).

100 Dumelie, J. G. & Jaffrey, S. R. Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. *Elife* **6**, doi:10.7554/eLife.28306 (2017).

101 Brazda, V., Haronikova, L., Liao, J. C. & Fojta, M. DNA and RNA quadruplex-binding proteins. *Int J Mol Sci* **15**, 17493-17517, doi:10.3390/ijms151017493 (2014).

102 Mishra, S. K., Tawani, A., Mishra, A. & Kumar, A. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep* **6**, 38144, doi:10.1038/srep38144 (2016).

103 Ryan, D. P. & Owen-Hughes, T. Snf2-family proteins: chromatin remodellers for any occasion. *Curr Opin Chem Biol* **15**, 649-656, doi:10.1016/j.cbpa.2011.07.022 (2011).

104 Castillo Bosch, P. *et al.* FANCJ promotes DNA synthesis through G-quadruplex structures. *EMBO J* **33**, 2521-2533, doi:10.15252/embj.201488663 (2014).

105 Wang, X. *et al.* Targeting of Polycomb Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines. *Mol Cell* **65**, 1056-1067 e1055, doi:10.1016/j.molcel.2017.02.003 (2017).

106 Long, Y. *et al.* Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *Elife* **6**, doi:10.7554/eLife.31558 (2017).

107 Kasinath, V. *et al.* Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science*, doi:10.1126/science.aar5700 (2018).

108 De, S. & Michor, F. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol* **18**, 950-955, doi:10.1038/nsmb.2089 (2011).

109 Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Non-canonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *bioRxiv*, doi:10.1101/146621 (2017).

110 Wang, G. & Vasquez, K. M. Effects of Replication and Transcription on DNA Structure-Related Genetic Instability. *Genes (Basel)* **8**, doi:10.3390/genes8010017 (2017).

111 McDonald, M. J. *et al.* Mutation at a distance caused by homopolymeric guanine repeats in Saccharomyces cerevisiae. *Sci Adv* **2**, e1501033, doi:10.1126/sciadv.1501033 (2016).

112 Lemmens, B., van Schendel, R. & Tijsterman, M. Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers. *Nat Commun* **6**, 8909, doi:10.1038/ncomms9909 (2015).

113 van Kregten, M. & Tijsterman, M. The repair of G-quadruplex-induced DNA damage. *Exp Cell Res* **329**, 178-183, doi:10.1016/j.yexcr.2014.08.038 (2014).

114 Choi, J. *et al.* A common intronic variant of PARP1 confers melanoma risk and mediates melanocyte growth via regulation of MITF. *Nat Genet* **49**, 1326-1335, doi:10.1038/ng.3927 (2017).

115 Casas-Selves, M. & Degregori, J. How cancer shapes evolution, and how evolution shapes cancer. *Evolution (N Y)* **4**, 624-634, doi:10.1007/s12052-011-0373-y (2011).

116 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

117 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

118 Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-954, doi:10.1038/nature00766 (2002).

119 Holderfield, M., Deuker, M. M., McCormick, F. & McMahon, M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat Rev Cancer* **14**, 455-467, doi:10.1038/nrc3760 (2014).

120 Prior, I. A., Lewis, P. D. & Mattos, C. A comprehensive survey of Ras mutations in cancer. *Cancer Res* **72**, 2457-2467, doi:10.1158/0008-5472.CAN-11-2612 (2012).

121 Su, F. *et al.* RAS mutations in cutaneous squamous-cell carcinomas in patients treated with BRAF inhibitors. *N Engl J Med* **366**, 207-215, doi:10.1056/NEJMoa1105358 (2012).

122 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

123 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).

124 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

125 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

126 Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**, 1258-1263, doi:10.1038/ng.3141 (2014).

127 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-1165, doi:10.1038/ng.3101 (2014).

128 Jiang, H., Ju, Z. & Rudolph, K. L. Telomere shortening and ageing. *Z Gerontol Geriatr* **40**, 314-324, doi:10.1007/s00391-007-0480-0 (2007).

129 Blackburn, E. H. Telomerase and Cancer: Kirk A. Landon--AACR prize for basic cancer research lecture. *Mol Cancer Res* **3**, 477-482, doi:10.1158/1541-7786.MCR-05-0147 (2005).

130 Barthel, F. P. *et al.* Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat Genet* **49**, 349-357, doi:10.1038/ng.3781 (2017).

131 Bell, R. J. *et al.* Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science* **348**, 1036-1039, doi:10.1126/science.aab0015 (2015).

132 Stern, J. L., Theodorescu, D., Vogelstein, B., Papadopoulos, N. & Cech, T. R. Mutation of the TERT promoter, switch to active chromatin, and monoallelic TERT expression in multiple cancers. *Genes Dev* **29**, 2219-2224, doi:10.1101/gad.269498.115 (2015).

133 Liu, R., Zhang, T., Zhu, G. & Xing, M. Regulation of mutant TERT by BRAF V600E/MAP kinase pathway through FOS/GABP in human cancer. *Nat Commun* **9**, 579, doi:10.1038/s41467-018-03033-1 (2018).

134 Zhang, T. *et al.* SDHD Promoter Mutations Ablate GABP Transcription Factor Binding in Melanoma. *Cancer Res* **77**, 1649-1661, doi:10.1158/0008-5472.CAN-16-0919 (2017).

135 Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22, doi:10.1016/j.ajhg.2017.06.005 (2017).

136 Price, A. L., Spencer, C. C. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci* **282**, 20151684, doi:10.1098/rspb.2015.1684 (2015).

137 Stormo, G. D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **11**, 751-760, doi:10.1038/nrg2845 (2010).

138 Dunham, W. H., Mullin, M. & Gingras, A. C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **12**, 1576-1590, doi:10.1002/pmic.201100523 (2012).

139 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355, doi:10.1038/nature19949 (2016).

140 Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637-643, doi:10.1038/nature04670 (2006).

141 Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712-723, doi:10.1016/j.cell.2015.09.053 (2015).

142 Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509, doi:10.1038/nature22366 (2017).

143 Liu, F., Rijkers, D. T., Post, H. & Heck, A. J. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat Methods* **12**, 1179-1184, doi:10.1038/nmeth.3603 (2015).

144 Chavez, J. D. *et al.* Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue. *Cell Syst* **6**, 136-141 e135, doi:10.1016/j.cels.2017.10.017 (2018).

145 Mittler, G., Butter, F. & Mann, M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res* **19**, 284-293, doi:10.1101/gr.081711.108 (2009).

146 Butter, F. *et al.* Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet* **8**, e1002982, doi:10.1371/journal.pgen.1002982 (2012).

147 Viturawong, T., Meissner, F., Butter, F. & Mann, M. A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation. *Cell Rep* **5**, 531-545, doi:10.1016/j.celrep.2013.09.022 (2013).

148 Kappei, D. *et al.* Phylointeractomics reconstructs functional evolution of protein binding. *Nat Commun* **8**, 14334, doi:10.1038/ncomms14334 (2017).

149 Butter, F., Scheibe, M., Morl, M. & Mann, M. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci U S A* **106**, 10626-10631, doi:10.1073/pnas.0812099106 (2009).

150 Klass, D. M. *et al.* Quantitative proteomic analysis reveals concurrent RNA-protein interactions and identifies new RNA-binding proteins in Saccharomyces cerevisiae. *Genome Res* **23**, 1028-1038, doi:10.1101/gr.153031.112 (2013).

151 Jazurek, M., Ciesiolka, A., Starega-Roslan, J., Bilinska, K. & Krzyzosiak, W. J. Identifying proteins that bind to specific RNAs - focus on simple repeat expansion diseases. *Nucleic Acids Res* **44**, 9050-9070, doi:10.1093/nar/gkw803 (2016).

152 Edupuganti, R. R. *et al.* N(6)-methyladenosine (m(6)A) recruits and repels proteins to regulate mRNA homeostasis. *Nat Struct Mol Biol* **24**, 870-878, doi:10.1038/nsmb.3462 (2017).

Chapter 2

Recurrent promoter mutations at *TERT* and *SDHD* in melanoma respectively recruit or abrogate GABP transcription factor binding

Modified from:
An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations.

Matthew M Makowski[*], Esther Willem[*], Jun Fang[*], Jiyeon Choi, Tongwu Zhang, Pascal WTC Jansen, Kevin M. Brown[#], Michiel Vermeulen[#].

Proteomics. 2016.

SDHD Promoter Mutations Ablate GABP Transcription Factor Binding in Melanoma.

Tongwu Zhang[*], Mai Xu[*], Matthew M Makowski[*], Christine Lee, Michael Kovacs, Jun Fang, Esther Willems, Jeffrey M Trent, Nicholas K Hayward, Michiel Vermeulen[#], Kevin M Brown[#].

Cancer Research. 2017.

Abstract

Recurrent somatic mutations in the human telomerase reverse transcriptase (TERT) promoter region, predominantly localized to two nucleotide positions, are highly prevalent in many cancer types. Indeed, aberrant telomerase reactivation in differentiated cells represents a major event in oncogenic transformation. Both mutations create novel consensus E26 transformation-specific (ETS) motifs and are associated with increased TERT expression. Here, we performed an unbiased proteome-wide survey of transcription factor binding at TERT promoter mutations in melanoma. We observed ELF1 binding at both mutations in vitro, yet we showed that increased recruitment of GABP is enabled by the spatial architecture of native and novel ETS motifs in the TERT promoter region. We characterized the dynamics of competitive binding between ELF1 and GABP and provided evidence for ELF1 exclusion by transcriptionally active GABP. Similarly, across cancer types, recurrent SDHD promoter mutations occur exclusively in melanomas, at a frequency of 4-5%. These mutations are predicted to disrupt consensus ETS-transcription factor binding sites and are correlated with both reduced SDHD gene expression and poor prognosis. Here, we found that expression of SDHD in melanoma correlated with the expression of multiple ETS-transcription factors, particularly in SDHD promoter wild-type samples. Consistent with the predicted loss of ETS-transcription factor binding, we observed that recurrent hotspot mutations resulted in decreased luciferase activity in reporter assays. Furthermore, we demonstrated specific GABPA and GABPB1 binding to probes containing the wild-type promoter sequences, with binding disrupted by the SDHD hotspot promoter mutations in both quantitative mass spectrometry and band-shift experiments. Finally, using siRNA-mediated knockdown across multiple melanoma cell lines, we determined that loss of GABPA resulted in reduced SDHD expression at both RNA and protein levels. These data are consistent with a key role for GABPA/B1 as the critical ETS-transcription factors deregulating TERT and SDHD expression in the context of highly recurrent promoter mutations in melanoma. Thus, we suggest a more careful search for other recurrent promoter mutations creating or disrupting GABPA consensus sequences is warranted.

Introduction

Compared to other human cancer types, cutaneous melanomas have a high mutation burden attributable to ultraviolet radiation (UVR) exposure[1,2]. The high number of mutations has complicated efforts to distinguish driver versus passenger mutations in large-scale sequencing studies[3-9]. To date, most genome-scale sequencing studies have relied heavily on analysis of exomes, identifying a spectrum of driver genes with recurrent protein-coding somatic mutations and establishing a generalized framework for the genomic classification of cutaneous melanoma[4-6,8,10]: *BRAF*-mutant, *RAS*-mutant, *NF1*-mutant, and "triple wild-type". Still, there is an emerging body of literature suggesting an important role for non-coding somatic mutations in melanoma development[4,5], including those found within the 5'-untranslated (UTR) regions of genes[4,11] and gene promoters[12-18]. Perhaps most notably, highly recurrent *TERT* promoter mutations that create consensus E26 transformation-specific transcription factor (ETS) binding motifs have been found in 50-85% of melanomas[12-14], as well as in the germline of two high-density melanoma families[12,15].

Telomerase reactivation, associated with senescence bypass and essentially unlimited proliferative capacity, is a classic hallmark of cancer [19,20]. However, because telomerase is typically silenced in differentiated cells, telomerase reactivation is thought to proceed through transcriptional mechanisms [21]. A number of factors have previously been implicated in TERT reactivation via various transcriptional pathways [22-24]. More recent work has begun to unravel the functional mechanisms by which TERT promoter mutations exert their oncogenic effect. For example, TERT promoter mutations were recently implicated in overcoming differentiation-associated transcriptional silencing of TERT expression, effectively extending telomere length and potentially contributing to tumorigenic immortalization [25,26]. Additionally, both mutations create novel E26 transformation-specific (ETS) transcription factor motifs within the TERT promoter region [27,28]. However, the exact factor or factors influencing TERT expression specifically via promoter mutation sites remained elusive until a recent study demonstrated that GABP, an ETS-family transcription factor, was a major functional interactor [29]. The identification of GABP as a mutation-specific interactor is particularly interesting, as multiple transcriptionally active binding modes for GABP have been previously been reported, including a heterodimer (GABPA/B) and a heterotetramer ($GABPA_2/B_2$) [30-34]. The heterotetrameric

binding mode of GABP was associated with TERT promoter mutation specific interactions. In particular, a precise spatial architecture between native ETS motifs and novel, mutation-specific ETS motifs was indicated as crucial for tetrameric GABP binding [29]. Very recently, another report indicated that mutation specific GABP binding is concurrent with a switch to active chromatin marks [35].

Recently, in an effort to identify similar hotspot mutations in gene promoters across multiple cancers sequenced as a part of The Cancer Genome Atlas (TCGA) project, Weinhold and colleagues identified recurrent mutations in the *SDHD* promoter exclusively in melanoma[18]. These mutations were associated with reduced levels of *SDHD* expression, as well as poor prognosis. These findings were replicated by Scholz and colleagues, who analyzed 451 melanomas and found that approximately 4% of samples harbored *SDHD* promoter hotspot mutations[36]. Consistent with the role of UVR in melanoma biology, *SDHD* promoter mutations occur primarily as C>T alterations in sun-exposed melanomas. The major mutations are located at chr.11:111,957,523 (TTCC>TTTC, C523T), chr.11:111,957,541 (TTCC>TTTC, C541T) and chr.11:111,957,544 (CTTCC>TTTCC, C544T)[18,36], within or adjacent to highly conserved TTCC motifs utilized by most ETS transcription factors[37]. While the ETS transcription factor family is one of the largest families of transcription factors, including more than 29 human genes[38], expression of *ELF1* was observed to be positively correlated with *SDHD* expression in TCGA samples without *SDHD* promoter mutation[18], suggesting a functional role for ELF1 in regulating *SDHD* transcription. Still, a direct role for ELF1 or other ETS family transcription factors in the regulation of *SDHD* in melanoma remains to be established.

Despite progress in elucidating the molecular function of TERT and SDHD promoter mutations, most studies thus far have targeted ETS-family factors. Complementary to targeted molecular studies, our group and others have previously established workflows for AP-MS/MS based identification of sequence specific protein-DNA binding on a proteome-wide scale [39,40]. Here, we perform an unbiased, proteome-wide survey of TERT promoter mutation specific transcription factor binding. We identify specific binding of multiple factors, including ELF1/2 and ETV6, and we confirm GABP as a direct, specific interactor in melanoma. We provide further characterization of the spatial architecture that promotes GABP binding at native and novel ETS motifs in

the TERT promoter. Additionally, we analyze competitive binding dynamics between ELF1 and GABP and propose a model in which multimeric GABP binding to a native and novel ETS motif excludes ELF1 and activates TERT expression. Similarly, we evaluate the incidence of *SDHD* promoter mutations in the three largest melanoma whole-genome and -exome datasets (TCGA/Broad/Yale). We then functionally assess the consequence of *SDHD* promoter mutations in melanoma using the same mass spectrometry based approach. We find that the ETS transcription factors GABPA and GABPB1 specifically bind to wild-type *SDHD* promoter sequences, with this binding disrupted by hotspot promoter mutations. Our findings here highlight the importance of transcription factors GABPA/B1 as key regulators of expression of select 'driver' genes in melanoma.

Materials and Methods

Cell culture and nuclear lysate extraction

Melanoma cell lines were grown in adherent culture in RPMI 1640 (Gibco) supplemented with 10% FBS, 200 mM HEPES (pH 7.9), 100 U/ml penicilin and 100 µg/ml streptomycin (Gibco).

Nuclear lysates were collected essentially as described previously [40]. Briefly, cells were incubated in hypotonic Buffer A (10 mM HEPES (pH 7.9), 1.5 mM MgCl2, 10 mM KCl and 0.15% NP40). Cells were then lysed by dounce homogenizer. Crude nuclei were collected by centrifugation and lysed in Buffer C (420 mM NaCl2, 20 mM HEPES (pH 7.9), 20% (v/v) glycerol, 2 mM MgCl2, 0.2 EDTA, 0.1% NP40, EDTA-free complete protease inhibitors (Roche), and 0.5 mM DTT) by rotation for one hour at 4C. Nuclear lysates were collected as the soluble fraction, snap frozen in liquid nitrogen, and stored at -80C.

DNA pulldown and on-bead sample preparation

Oligo baits were ordered via custom synthesis from Biomers or IDT with 5'-biotinylation of the forward strand. Oligos were combined with 1.5X molar excess of the reverse strand in 2X annealing buffer (20 mM TRIS, pH 8.0, 100 mM NaCl and 2 mM EDTA) and denatured at 98C for 10 minutes. Oligos were allowed to anneal by cooling to room temperature overnight and subsequently stored at -20C. For each pulldown reaction, 20 µl bead slurry (10

µl beads) of Streptavidin-Sepharose beads (GE Healthcare) were used. Each pulldown was performed in duplicate for outlier calling. Label swapping was performed between replicates to eliminate labeling bias. All pulldowns from each labeling reaction (forward and reverse) were performed simultaneously. Beads were washed once with 0.1% NP40 in 1X PBS and once with DNA binding buffer (DBB: 1M NaCl, 0.05% NP40, 10 mM TRIS, pH 8.0 and 1 mM ETDA). Annealed oligo (500pmol) was diluted in 600 µl DBB final volume and rotated for 30 min at 4C. Subsequent steps were all carried out at 4C. Beads with immobilized oligonucleotides were washed once with DBB and twice with protein binding buffer (PBB: 150 mM NaCl, 0.25% NP40, 50 mM TRIS, pH 8.0, EDTA-free complete protease inhibitors, and 1 mM DTT). Nuclear extracts (500ug) and 10 µg of competitor DNA (5 µg poly-dIdC, 5 µg poly-dAdT) were added to beads in a 600 µL final volume. For all AP-MS/MS analyses, nuclear extract from UACC903 cells (TERT C228T-positive) was used. Beads were incubated for 90 min on a rotation wheel at 4°C. The beads were then washed three times with PBB and two times with 1X PBS. All supernatant was carefully removed with a syringe. The proteins were reduced in elution buffer (2 M urea, 10 mM DTT, and 100 mM ammonium bicarbonate) for 20 minutes with shaking at room temperature. Samples were alkylated by addition of 50 mM iodoacetamide (IAA) in the dark with shaking at room temperature for 10 minutes. Proteins were then subjected to on-bead trypsin digestion (0.25 µg) for 2 hours at room temperature plus shaking. The supernatant was transferred to a new tube and digested with an additional 0.1 µg trypsin overnight.

On-StageTip Dimethyl Labeling

Tryptic peptides were purified on C18 stage-tips (without acidification) as described previously [41]. Buffer A was 0.1% formic acid and Buffer B was 80% acetonitrile and 0.1% formic acid. On-StageTip dimethyl labeling was performed as described previously [42]. Briefly, 300 µl of labeling reagent (16.2 µl 37% CH2O (light) or 30.0 µl 20% CD2O (medium) plus 6 mg sodium cyanoborohydride in 3mL of labeling buffer [10 mM NaH2PO4, 35 mM Na2HPO4]) was applied to the StageTip and spun through at 2200g for 10 min. StageTips were then washed once with 100uL of Buffer A and stored at 4C for MS analysis.

Mass spectrometry analysis

Labeled samples were eluted from the StageTips with 30 µl of Buffer B while combining the respective light and medium labeled pairs into the same tube. Acetonitrile was evaporated by SpeedVac centrifuge at room temperature. After resuspension with 7 µl of Buffer A, 5 µl of sample was loaded onto a 30 cm column (heated at 40°C) packed in-house with 1.8 µm Reprosil-Pur C18-AQ (Dr Maisch). The peptides were eluted from the column using a gradient from 7 to 32% Buffer B in Buffer A over 120 minutes at flow rate 250 nl/min using an Easy-nLC 1000 (Thermo Fisher Scientific).

TERT C228T and C250T and all SDHD mutation samples were eluted and sprayed directly into a Thermo Fisher QExactive mass spectrometer. The mass spectrometer was operated in top10 data-dependent acquisition mode. Target values for full MS were set to 3e6 AGC target and a maximum injection time of 20 ms. Full MS were recorded at a resolution of 70,000 over a scan range of 300-1650 m/z. Target values for MS/MS were set at 1e5 AGC target with a maximum injection time of 120 ms. The MS/MS spectra were recorded at a resolution of 17,500. The isolation width was set to 3.0 m/z, the collision energy to NCE=25, and the intensity threshold to 8.3e2. Dynamic exclusion was enabled for 20 s. Peptides with single or unknown charge state were excluded for MS/MS analysis.

TERT C228T+ETS samples were eluted and sprayed directly into a Thermo Fisher Orbitrap Fusion Tribrid mass spectrometer. Target values for full MS were set to 4e5 AGC target and a maximum injection time of 50 ms. Full MS were recorded at a resolution of 120,000 at a scan range of 400-1500 m/z. Most intense precursors with a charge state between 2 and 7 were selected for MS/MS analysis, with an intensity threshold of 5000 and dynamic exclusion for 60 s. Target values for MS/MS were set at 1e4 AGC target with a maximum injection time of 35 ms. Ion trap scan rate was set to 'Rapid', with an isolation width of 1.6 m/z and collision energy of 35%.

Data analysis

Raw MS spectra were analyzed using MaxQuant software (version 1.5.1.0) with standard settings [43,44]. For dimethyl labelled samples, the respective built in N-terminal and lysine modification for dimethyl labeling was specified under "Labels". Carbamidomethylation was specified as a fixed modification

on cysteines. N-terminal acetylation and methionine oxidation were allowed as variable modifications. Trypsin was selected as specific enzyme, and two missed cleavages were allowed. Data was searched against the human UniProt database (fasta file downloaded 2014.09.03) using the integrated search engine. The search was performed with a mass tolerance of 4.5 ppm mass accuracy for the precursor ion and 20 ppm for fragment ions. Peptides and proteins were both accepted at an FDR of 0.01. For quantification, at least two ratio counts were required. Protein identifications and calculated ratios are included as the proteinGroups output file from MaxQuant analysis. Plots were generated with Python essentially as described previously [45]. Briefly, protein identifications were filtered for contaminants and reverse hits. Proteins groups were required to have two identified peptides, of which at least one was unique, to be considered as identified. The required outlier significance was 3.0 IQRs (inter-quartile range) for both forward and reverse experiments.

Band-shift analysis of recombinant protein-DNA interactions

Band-shift experiments were performed by incubating 20 fmol of biotin labeled double stranded oligos with recombinant human ELF1 protein (Origene, TP760629) or recombinant GABPA/B (Abnova, GABPA: H00002551-P01, GABPB: H00002553-P01) in a total volume of 20 µl of protein binding buffer (PBB: 150 mM NaCl, 0.25% NP40, 50 mM TRIS, pH 8.0, and 1 mM DTT) for 30 minutes. For GABP experiments, GABPA and GABPB were mixed at equimolar concentrations for 20 minutes at room temperature prior to addition of the oligo, In GABP experiments, the molecular weights listed in the figures refers to the total molecular weight of protein used (GABPA/B combined). The resulting protein complexes were resolved on 4-20% TBE gels (Biorad) in a Mini-PROTEAN tetra cell (Biorad) at 100V for approximately 3 hours in 1X TBE. Samples were transferred onto a nylon membrane (Biodyne) in a Trans-Blot Turbo Transfer semi-dry transfer system (Biorad) at 400 mA for 10 minutes. Membranes were UV cross-linked and oligos were detected using streptavidin-HRP conjugate and a chemiluminescent substrate (Chemiluminescent Nucleic Acid Detection Module, Pierce).

Western blotting

For TERT mutation western blot analysis, DNA pulldowns were performed as described above. After the final PBS wash, samples were resuspended in MilliQ water plus 1X sample buffer and boiled at 95C for ten minutes. Samples were then resolved on a poly-acrylamide gel, transferred to a nitrocellulose membrane, blocked with 5% milk, and incubated with primary antibody at 4C overnight (ELF1: Santa Cruz, sc-631; GABPA: Santa Cruz, sc-22810). Samples were then incubated with HRP-conjugated secondary antibody (Dako) for one hour at room temperature, and imaged using an ECL Western Blotting Substrate (Pierce).

For SDHD mutation western blot analysis, total cell lysates were generated with RIPA (Thermo Scientific, Pittsburgh, PA) and subjected to water bath sonication. Samples were resolved by 4-12% Bis-Tris ready gel (Invitrogen) electrophoresis. The primary antibodies used were rabbit anti-SDHD (ab189945, Abcam), rabbit anti-GABPA (ABE1047, Millipore), mouse anti-GABPB1 (sc271571, Santa Cruz Biotechnology), and mouse anti-β-actin (A5316, Sigma-Aldrich).

RNA and genomic DNA extraction

RNA was extracted using an RNeasy Plus Mini Kit (Qiagen). Genomic DNA was isolated using the ZR genomic DNA™ kit (D3050, ZYMO Research) and assessed by Nanodrop 8000 (Thermo Scientific).

Long PCR and sequencing of TA clones

PCR was carried out using genomic DNAs from 7 UACC melanoma cell lines as indicated in Figure 4A. The PCR primers used were: GGGCCGCAGCTGCTCCTTGTCG and CAGGCCGGGCTCCCAGTGG. PCR conditions were 98C for 10 minutes to initially denature, followed by 38 cycles of: 98C for 30s, 60C for 30s, and 72C for 90s, with a 7-min final extension at the end. The PCR products were cloned into TA vector (Life Technologies). Single colonies of bacteria were selected and PCR sequenced in 96-well plate format. The PCR reaction was performed in the same manner as above. Sanger sequencing was used to determine the sequences of the PCR product.

Real-time quantitative PCR and allele specific TERT expression

For TERT mutations, gene expression levels were quantified by quantitative real-time PCR using TaqMan assays for TERT (Hs00972656_m1), ELF1 (Hs01111177_m1), ELF2 (Hs00959420_g1), GABPA (Hs01022016_m1), and GAPDH (cat#4333764) from Life Technologies. siRNA knockdown experiments were performed using siRNAs purchased from Dharmacon targeting GABPA (D-001810-01-05), ELF1 (L-012669-00-0005), and ELF2 (L-012754-00-0005). A non-targeting scrambled siRNA was used as the control (D-001810-01-05). Gene expression levels of TERT, ELF1 and ELF2 were normalized to GAPDH. Allele specific TERT expression was determined using an allelic discrimination TaqMan assay for rs2736098 (assay C_26414916_20, Life Technologies), and the gene expression of each allele of TERT was also normalized to the gene expression of GAPDH. Each experiment was performed in triplicate and repeated three times.

For SDHD mutations, pools of 4 siRNAs each respectively targeting one transcription factor gene (*ELF1*, *PRDM1*, *IRF4*, *GABPA* and *GABPB1*), as well as a non-specific control siRNA, were purchased from GE Dharmacon. siRNAs were transfected into human melanoma cell lines using Lipofectamine RNAiMAX (Life Technologies). At day 2-7 following transfection, total RNA was extracted from cells using RNeasy Minikit (Qiagen), followed by cDNA synthesis (iScript™ cDNA Synthesis Kit; BioRad, Hercules, CA). Quantitative real-time PCR was performed using Taqman assays (Invitrogen, Carlsbad, CA). *GAPDH* served as an internal control. For western blot analysis, total cell lysates were generated with RIPA (Thermo Scientific, Pittsburgh, PA) and subjected to water bath sonication. Samples were resolved by 4-12% Bis-Tris ready gel (Invitrogen) electrophoresis. The primary antibodies used were rabbit anti-SDHD (ab189945, Abcam), rabbit anti-GABPA (ABE1047, Millipore), mouse anti-GABPB1 (sc271571, Santa Cruz Biotechnology), and mouse anti-β-actin (A5316, Sigma-Aldrich).

Melanoma sequencing datasets

Melanoma whole-exome sequencing (WES) or whole-genome sequencing (WGS) datasets (BAM files) were downloaded from CGHub (https://cghub.ucsc.edu) for TCGA SKCM samples (n=470; http://cancergenome.nih.gov)[4], or dbGaP (http://www.ncbi.nlm.nih.gov/gap) for Broad Institute[3] and Yale[7,8]

datasets (Broad, n=122, phs000452.v1.p1; Yale, n=213, phs000933.v1.p1). TCGA mRNA expression data was downloaded from cBioPortal (http://www.cbioportal.org; RNA Seq V2 RSEM)[46]. We additionally collected exome sequencing data from previously published[47] (n=44; European Nucleotide Archive, PRJEB11984) and 55 additional melanoma cell lines (obtained from the University of Arizona Cancer Center in 2007; UACC). Cell lines were initially characterized via Sanger sequencing of 10 melanoma driver genes, re-authenticated via exome sequencing (cells were simultaneously microsatellite profiled at that time, 2012; AmpFLSTR Identifiler, ThermoFisher) and re-authenticated via microsatellite profiling immediately prior to functional experiments described below (2016). Reads for all samples were aligned to the human genome (hg19) with the Burrows-Wheeler Aligner (BWA 0.6.2)[48] and processed with the Genome Analysis Toolkit (GATK 2.3)[49,50] including local realignment and base quality recalibration.

SDHD promoter mutation identification

We applied a pipeline utilizing bam files to all WGS/WES data in this study. *SDHD* promoter regions were defined as being 0-500 bp upstream in RefGene (http://genome.ucsc.edu/cgiDbin/hgTables). bam-readcount was used to count bases[51], and the following criteria was applied to identify recurrent promoter mutations: (1) mutation was only found in tumors; (2) sequencing depth for each mutation location was greater than 6; (3) alternative base count was greater than 2; (4) average mapping quality of each mutation location was greater than 20; (5) average base quality of each mutation location was greater than 20. Cell line mutations were validated by Sanger sequencing on a 3730xl DNA analyzer (ABI) (primers, F: TCCGCCATTGTTCGCCTC and R: CTCCAGAGAACCGCCATCTC), with forward and reverse traces analyzed using Mutation Surveyor (SoftGenetics). Expression correlation and statistical tests were performed using R (https://www.r-project.org).

Motif analysis

Prediction of mutation effects on transcription factor binding sites was performed using the motifbreakR package[52] and a comprehensive collection of human transcription factor binding sites models (HOCOMOCO)[53]. We applied

the information content algorithm as the method, and used a threshold of 0.0001 as the maximum *P*-value for a transcription binding site match in motifbreakR.

SDHD promoter luciferase reporter assays

Five luciferase constructs for SDHD mutations (wild-type, C523T, C524T, C541T and C544T) were generated to containing 163bp of the genomic sequence surrounding *SDHD* promoter mutations (Chr11: 111,957,437-111,957,599). The fragment was PCR-amplified (primers, F: CTGAACTctcgagCTCCGCCATTGTTCGCCTC, R: GTCACTGTagatctACCCGGAACCACTTAGGCGAC) from genomic DNA purified from cells harboring wild-type and mutant *SDHD* promoter, sub-cloned into pGL4.23[luc2/minP] (Promega) luciferase vector, and constructs sequence-verified. Constructs were co-transfected with pGL4.74 (renilla luciferase) into human melanoma cell lines using Lipofectamine 2000 (Life Technologies). Cells were collected 24 hrs after transfection and luciferase activity was measured using the Dual-Luciferase reporter system (Promega) on GLOMAX Multi Detection System (Promega).

Results

ELF1 binds specifically to both TERT promoter mutations

We used an AP-MS/MS based workflow to identify the proteome-wide interactomes of both TERT promoter mutations. Oligo baits were designed to encompass both mutation sites concurrently, with one mutation per oligo (Fig 1A). For AP-MS/MS analysis, we used UACC903 metastatic melanoma derived cell lines, characterized as C228T-positive. We identified ELF1 and ELF2 as specific interactors at both the C228T and C250T mutation sites and ETV6 as a specific interactor at the C250T mutation site (Fig. 1B, 1E). We confirmed the specificity and robustness of the ELF1 interaction via bandshift with recombinant protein (Fig. 1C, 1F). JASPAR *in silico* motif prediction agrees with mutation specific binding for ELF1 and GABPA (Fig. 1D, 1G) [54,55]. However, we did not observe a specific GABP interaction with either TERT promoter mutation by MS analysis. Intriguingly, band-shift experiments using pure, recombinant GABP revealed a subtle preference for the C228T and C250T mutations over the wild-type sequence (Fig. 1C, 1F). This potentially

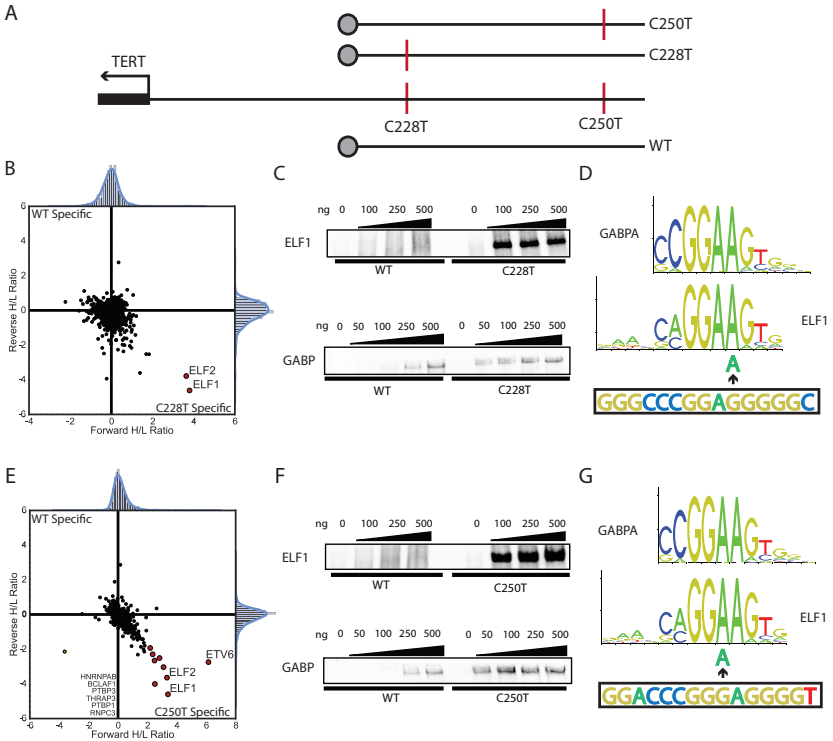


Figure 1. AP-MS/MS identifies TERT promoter mutation specific interactors

A Custom oligos were 5' biotinylated and designed to cover both the C228T and C250T mutation sites. Oligos are referred to by mutation as shown in the diagram.

B AP-MS/MS analysis of the C228T TERT promoter mutation interactors. Interactors with a ratio of at least 3 IQR (inter-quartile range) in both experiments are colored in red. Ratios are shown after log2 transformation. Labels were swapped between replicates to avoid labeling bias, hence specific interactors showing a high ratio in one experiment, and a low ratio in the other. Outlier proteins not observed consistently across experiments are noted at the bottom of the chart.

C Band-shift experiments with the C228T TERT promoter mutation oligo. Recombinant human protein was incubated with annealed oligo and resolved on a TBE polyacrylamide gel. Band-shift experiments for each protein-oligo combination were resolved on the same gel at the same exposure.

D JASPAR motif prediction agrees with mutation specific ELF1 and GABPA binding at C228T [54,55]. The underlying wild-type sequence surrounding the C228T mutation is shown beneath the JASPAR motifs, and the mutation is indicated by arrow.

E AP-MS/MS analysis of C250T TERT promoter mutation interactors.

F Band-shift experiments with the C250T TERT promoter mutation oligo and recombinant human protein.

G JASPAR motif prediction agrees with mutation specific ELF1 and GABPA binding at C250T. The underlying wild-type sequence surrounding the C250T mutation is indicated below the JASPAR motifs as in E).

indicates that the stabilizing effect of a single mutation (and thus a single canonical motif) is sufficient to relatively increase recombinant GABP binding *in vitro*. Indeed, comparison between previous GABP electrophoresis analysis and TERT promoter bandshift experiments suggests that the low-mobility complex we observed is indeed the GABPA/B heterotetramer [56]. However, GABP concentrations in nuclear lysates are likely far lower than those used in bandshift experiments, which explains why GABP was not identified by AP-MS/MS Our data thus suggests that ELF1 is a significant, specific interactor of single ETS motifs created by recurrent mutations in the TERT promoter. However, TERT promoter mutation-specific GABP binding at single novel ETS motifs, while present, appears to be of lower affinity.

GABP binding at novel and native ETS motifs excludes ELF1 and activates TERT expression

To resolve the discrepancy between ELF1 binding at TERT promoter mutations *in vitro* and its lack of transcriptional effect *in vivo*, we again performed AP-MS/MS analysis of TERT promoter mutations, utilizing the C228T mutation with two upstream native ETS motifs as a representative case (Fig. 2A)[29]. This oligo design facilitated study of combinatorial binding with novel and native ETS motifs. Indeed, recent work indicates that in the case of 250T, GABP still binds using ETS-195 and ETS-200 instead of upstream ETS motifs [29]. These native motifs (ETS-195, ETS-200) represent a different spatial architecture with either C228T or C250T compared to C228T and C250T with each other, which likely has implications for the mutual exclusivity of C228T and C250T mutations *in vivo*. For C228T+ETS, we observed robust and specific binding of GABP by MS analysis, and we could confirm the relative specificity of this binding by band-shift assay (Fig 2B, 2C).

We observed increased GABP binding concurrent with a reduction in specific ELF1 and ELF2 binding; indeed, neither reached significance at thresholds chosen in our previous analysis. By competition band-shift assay, we were able to directly compare ELF1 and GABP binding using both the C228T and the C228T+ETS oligos. With the C228T oligo, we observed relatively higher ELF1 binding compared with GABP in equimolar conditions (Fig 2D). However, with the C228T+ETS oligo, we observed increased binding for GABP in equimolar conditions (Fig. 2E). This supports our MS analysis

indicating that heterotetrameric GABP binding excludes ELF1 by occupying both the native and the novel ETS motifs present in this sequence. This further indicates that the stabilizing effect of a single TERT promoter mutation, in conjunction with binding at native ETS motifs present in C228T+ETS recruits GABP binding more robustly than does C228T or C250T. Furthermore, by western blot we were able to support our MS and band-shift data, showing increased ELF1 binding over wild-type at the C228T, C250T, and to a lesser extent the C228T+ETS sequences, and increased GABP binding over wild-type at the C250T and C228T+ETS sequences but not the C228T sequence (Fig. 2F). This strongly indicates that the GABP binding observed at C228T+ETS critically depends on the native ETS motif, as no specific binding is seen at C228T. Also, preferential binding at C250T but not C228T, suggests GABP binding at 228T versus C250T might operate under different kinetics [29,57].

To study the relationship between GABP, ELF1, and ELF2 binding and TERT expression *in vivo*, we used a PCR-based Sanger sequencing approach to identify seven melanoma cell lines with both a heterozygous germline SNP variant (rs2736098) and a haplotype-phased heterozygous TERT promoter mutation. In these cell lines, we observed mono-allelic TERT expression in which the expressed SNP allele always correlated with a phased TERT promoter mutation (Fig. 3A). Then, we used siRNA-mediated knockdown to discern the direct transcriptional effects of ELF1/2 and GABPA on TERT expression. In two separate cell lines, we observed a substantial reduction in TERT expression upon GABPA knockdown but only minimally reduced TERT expression upon combined ELF1/2 knockdown (Fig. 3B, 3C). Thus, this data strongly indicates that TERT reactivation proceeds via mono-allelic TERT expression exclusive to the promoter mutation phased chromosome. Upon displacement of minimally activating ELF1 or ELF2, GABP binding induces a robust transcriptional response driving TERT expression. This study suggests a model where binding of heterotetrameric, transcriptionally active GABP at novel and native ETS motifs effectively precludes ELF1 occupancy at the TERT promoter mutations (Fig. 4).

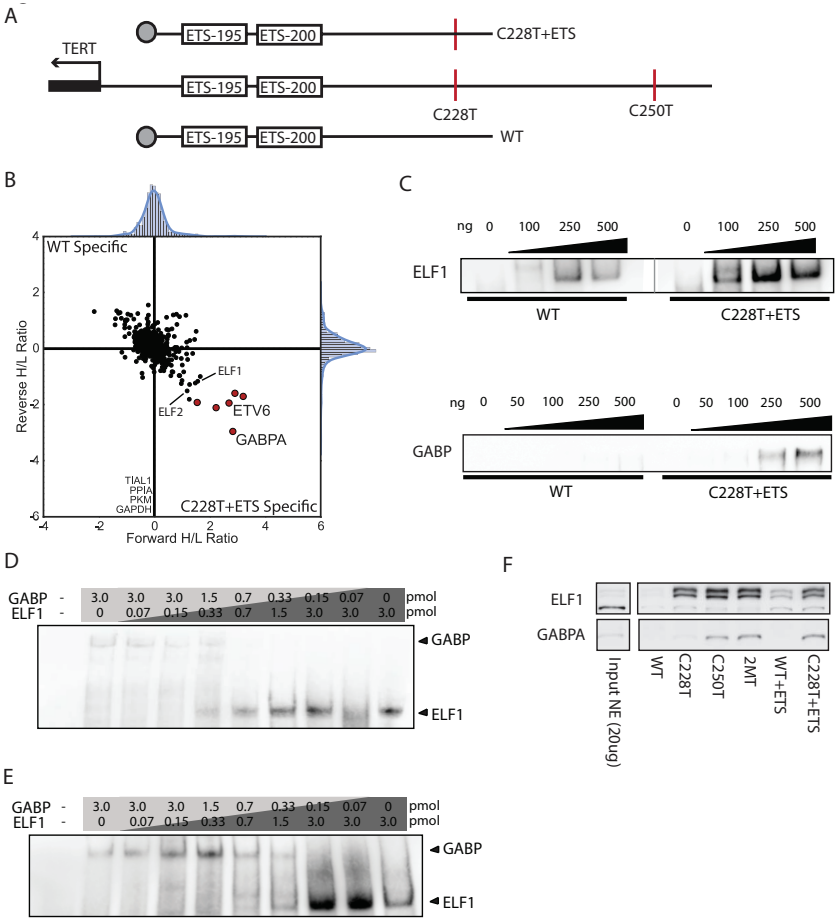


Figure 2. Heterotetrameric GABP excludes ELF1 from TERT promoter mutation binding

A Oligos were designed as before, but encompassing the C228T mutation plus two additional upstream ETS motifs.

B AP-MS/MS analysis of the C228T+ETS interactors. ELF1 and ELF2 were not significant at a 3 IQR cutoff, but were noted specifically for comparison to GABPA and previous binding ratios.

C Band-shift experiments with the C228T+ETS TERT promoter mutation oligo. The grey line in the ELF1 band-shift indicates re-positioning of lanes; all lanes were run on the same gel, and exposures were kept uniform.

D Competitive binding experiments between recombinant human ELF1 and GABP at the C228T TERT promoter mutation oligo.

E Competitive binding experiments between recombinant human ELF1 and GABP at the C228T+ETS TERT promoter mutation oligo.

F Western blot analysis of GABPA and ELF1 for each oligo used in this study. All lanes are taken from the same gel at the same exposure.

A

Cell line	Expression T-allele/C-allele	Expressed allele	TERT Promoter Mutation Strand: T-allele	TERT Promoter Mutation Strand: C-allele
UACC91	>100	T	+	-
UACC383	>100	T	+	-
UACC457	<0.01	C	-	+
UACC502	<0.01	C	-	+
UACC1592	>100	T	+	-
UACC1940	>100	T	+	-
UACC2331	<0.01	C	-	+

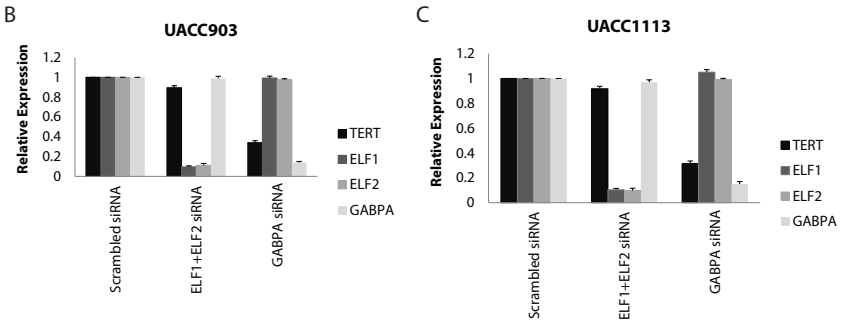


Figure 3. GABP activates mono-allelic TERT expression via promoter mutations

A TERT expression is mono-allelic and correlates with promoter mutation status. Mono-allelic TERT expression was assayed at rs2736098. Expression of only one allele was observed. The expressed allele correlated with a phased TERT promoter mutation as observed by PCR-based Sanger sequencing.

B In the UACC903 melanoma cell line, GABPA knockdown substantially reduces TERT expression, while combined ELF1/2 knockdown minimally reduces TERT expression. Experiments were performed in triplicate and repeated three times. Error bars represent standard error of the mean.

C In the UACC1113 melanoma cell line, GABPA knockdown substantially reduces TERT expression, while combined ELF1/2 knockdown minimally reduces TERT expression.

Identification of SDHD promoter mutations in multiple melanoma sequencing studies

Like TERT, though at much lower frequency, SDHD promoter mutations have previously been reported to show an ETS-family specific mutational signature in melanoma. In order to investigate *SDHD* promoter mutations in publicly available melanoma sequencing data, we downloaded WES data for the three largest melanoma sequencing studies (TCGA SKCM=470, Broad=122 and Yale=213), high coverage WGS data for TCGA SKCM data (n=40) from CGHub and dbGaP, and supplemented with WES data generated from a panel of melanoma cell lines (n=99). This sample size here (n=904) was considerably larger than the original *SDHD* promoter mutation study reported by Weinhold and colleagues (17 whole-genomes, 128 whole-exomes; TCGA). We searched

this larger dataset for somatic mutations within the *SDHD* promoter and 5'UTR (hg19 Chr11:111,957,493-111,957,631). Within the TCGA dataset, five recurrent mutations were identified, including the three (C523T, C541T, C544T) reported by Weinhold and colleagues, as well as two additional mutations (C532A, C548T; Fig. 5). Analysis of the larger combined dataset identified a total of ten mutations observed in more than one melanoma sample, with all but one (C532A) found in multiple datasets; mutations observed in cell lines (seven mutations in eight cell lines) were all confirmed via Sanger sequencing (Supplementary Fig. S1 and data not shown). The overall frequency of all *SDHD* promoter mutations was 5% (46/904, Supplementary Table S1), consistent with previous reports. The most frequently observed mutations (C523T, C544T, C541T, C524T) all are predicted to disrupt consensus ETS transcription factor binding sites, and were thus chosen to further investigate the mutational consequence in melanoma.

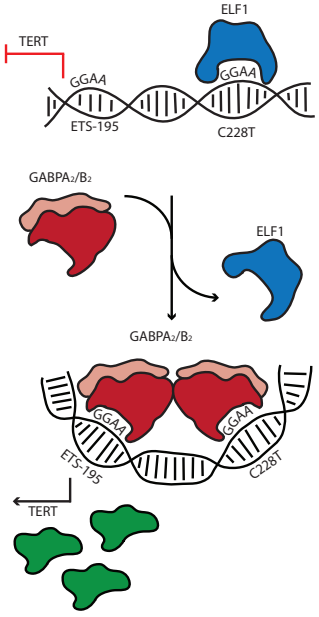


Figure 4. A model for ELF1 exclusion by heterotetrameric GABP at TERT promoter mutations

ELF1 binds *in vitro* at novel ETS motifs created by TERT promoter mutations. However, GABP binds with high affinity at novel and native ETS motifs and excludes ELF1 from binding.

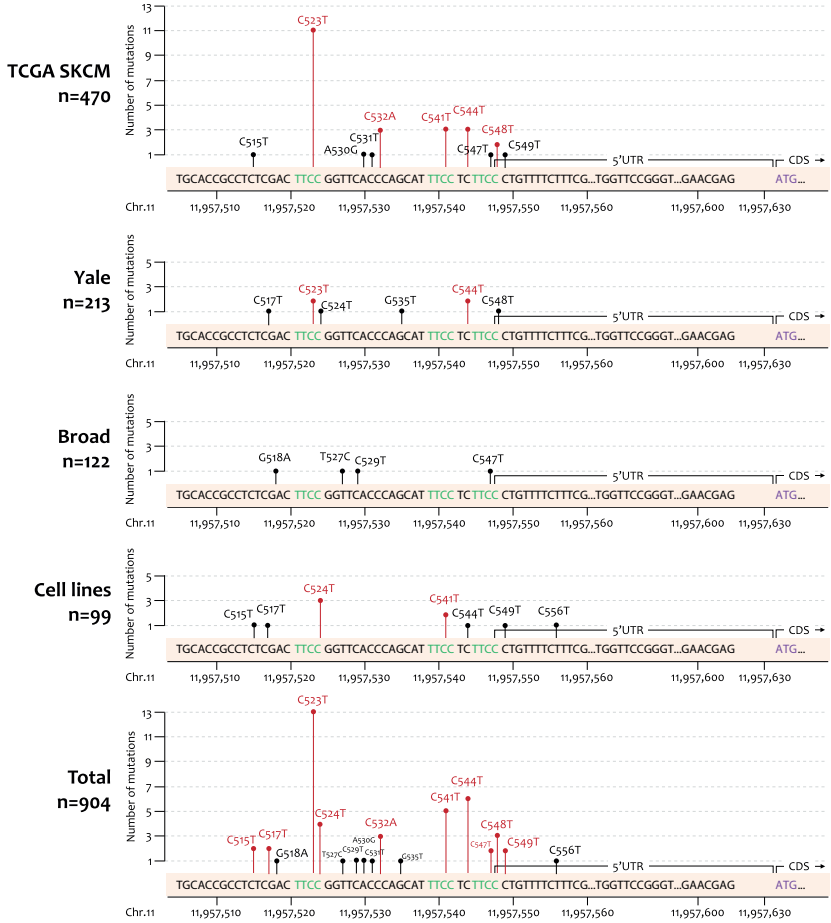


Figure 5. SDHD promoter mutation identification in multiple melanoma tumors datasets and cell lines.

The frequencies of SDHD promoter mutations are identified from whole-genome and –exome sequencing data of 470 TCGA SKCM, 213 Yale, and 122 Broad melanoma tumors, as well as 99 melanoma cell lines. Most recurrent mutations (colored in red) within each dataset are located within or adjacent to three consensus ETS transcription binding motifs ("TTCC", colored in green).

Allele-specific gene regulatory potential for SDHD hotspot promoter mutations

To investigate the correlation between *SDHD* promoter mutations and *SDHD* mRNA expression in melanoma, we downloaded mRNA expression data for

TCGA SKCM samples from cBioPortal (RNA-Seq V2 RSEM, n=470). Several of the hotspot *SDHD* promoter mutations found in this dataset are predicted to disrupt consensus ETS transcription factor binding sites, and thus might be expected to result in reduced *SDHD* gene expression. Considering samples that are copy-neutral at the *SDHD* locus, we observed significantly lower *SDHD* expression in those harboring the most common (C523T) mutation relative to wild-type samples (one-tailed student's t-test, $P = 5.97 \times 10^{-4}$, Benjamini & Hochberg adjusted $P = 0.002$, Fig. 2A). Despite small sample numbers, we also observed decreased expression for C541T ($P = 0.013$; adjusted $P = 0.026$), however differences for C548T ($P = 0.050$; adjusted $P = 0.066$) and C544T (adjusted and unadjusted $P = 0.050$) were not significant after adjusting for multiple testing (C532A was unassessable, n=1). Considering all samples without regard to copy number, we also observed significantly lower expression in samples with the C523T mutation relative to wild-type (one-tailed student's t-test, $P = 0.014$, adjusted $P = 0.041$; Supplementary Fig. S2).

To further evaluate the functional consequences of *SDHD* promoter mutations on *SDHD* expression, we next performed luciferase reporter assays. We cloned the four most common hotspot promoter mutations occurring within or immediately adjacent to conserved ETS motifs ("TTCC"; C523T, C524T, C541T, and C544T) and wild-type promoter sequence into a luciferase vector. We tested these constructs in multiple cell lines that varied in terms of both *SDHD* promoter mutation status (C021, C541T and C517T; C077, C541T and C544T; UACC1113, WT; and UACC903, WT) and relative endogenous expression of *SDHD* (higher expression in C021 and UACC1113; relatively low levels in C077 and UACC903). Compared to the promoterless vector, the wild-type vectors exhibited strongly increased luciferase activity in all four cell lines (Fig. 6B and Supplementary Fig. S3). C523T and C524T resulted in significant reductions of reporter expression across all four cell lines (two-tailed student's t-test *P*-value ranged from 4.35×10^{-11} to 9.37×10^{-07} and 4.17×10^{-10} to 3.60×10^{-04}, respectively. Fig. 6B). The C544T mutation, which occurs directly adjacent to a "TTCC" sequence (CTTCC>TTTCC), resulted in a significant reduction of reporter expression in three cell lines (two-tailed student's t-test: UACC1113, $P = 0.043$; C021, $P = 6.30 \times 10^{-06}$; UACC903, $P = 0.01$), whereas C541T showed significant reductions in two of the four cell lines tested (two-tailed student's t-test: UACC1113, $P = 0.0053$; C021, $P = 2.86 \times 10^{-05}$).

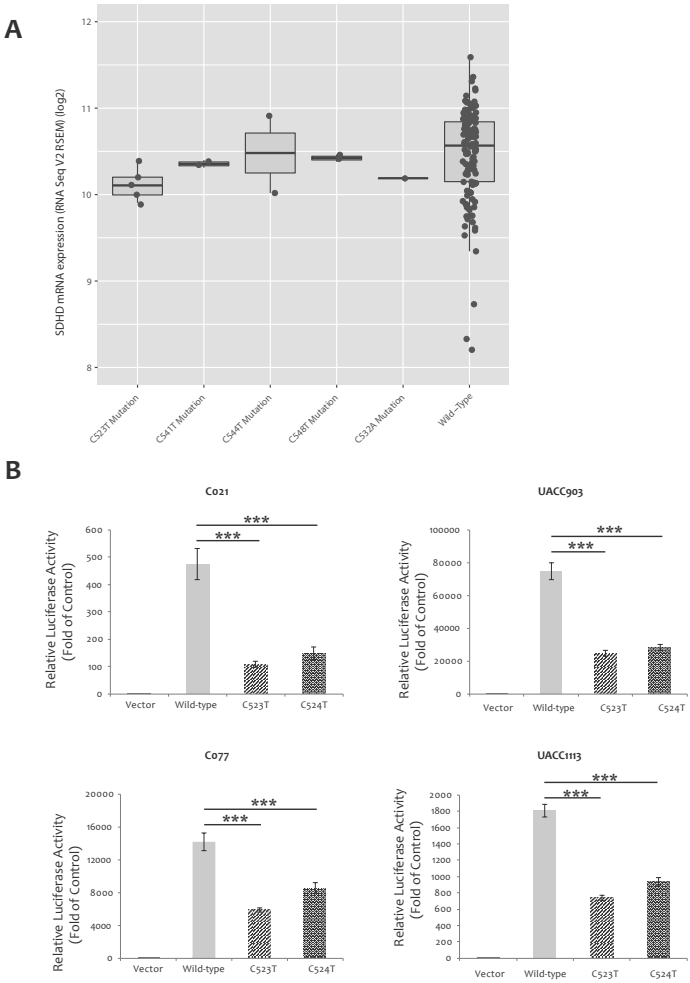


Figure 6. SDHD expression difference in SDHD copy-neutral melanomas harboring promoter mutations compared to wild-type samples.

A SDHD mRNA expression is significantly decreased in TCGA SKCM samples harboring the C523T and C541T mutations relative to wild-type samples.

B SDHD promoter activity is significantly decreased by SDHD hotspot mutation (C523T and C524T). A 163-bp fragment from the wild-type SDHD promoter sequence surrounding hotspot mutations significantly enhance luciferase reporter expression relative to vector control, whereas the same fragment containing hotspot mutations decrease promoter activity relative to the wild-type sequence. Fold change over minimal promoter control (vector only) is plotted as relative luciferase activity. The experiment was performed four times with triplicates for each. Stars denote significant differences in luciferase activity by two-tailed student's t-test (*: P-value <0.05; **: P-value <0.01; ***: P-value <0.001).

These data are consistent with an interpretation that these mutations result in reduced levels of *SDHD* gene expression, while suggesting a potentially larger effect for the more commonly observed C523T mutation as well as the adjacent C524T mutation.

Effects of SDHD promoter mutations on ETS transcription factor binding

To predict mutational effects on transcription factor binding sites in the *SDHD* promoter, we performed motif analyses for the four most common recurrent mutations that occurred within or directly adjacent to a "TTCC" motif: C523T, C524T, C541T and C544T. As expected, the C523T mutation was predicted to disrupt multiple transcription factor binding sites, 13/16 of which were ETS transcription factor binding sties (TTCC >TTTC). We observed the same effect for the C524T mutation. These mutations were predicted to have the strongest effect on GABPA binding (Altscore-Refscore -2.0 and *P*-value increased from 6.68×10^{-6} to 4.25×10^{-3} for both C523T and C524T, Fig. 7 and Supplementary Table S2), as well as a weaker effect on binding of ELF1 (Altscore-Refscore -1.88 and *P*-value increased from 3.35×10^{-6} to 7.43×10^{-4} for both C523T and C524T). In contrast, the C541T mutation both created and altered consensus motifs with strongest effect on PRDM1 binding (Supplementary Fig. S4A), while C544T was predicted to only create new motifs with strongest effect on IRF4 binding (Supplementary Fig. S4B).

We subsequently used the TCGA SKCM gene expression dataset to evaluate the correlation between gene expression of predicted ETS transcription factors and *SDHD*, in both *SDHD* wild-type samples and samples bearing *SDHD* promoter mutations. Of the 13 ETS transcription factors for which binding sites are predicted to be altered by the C523T mutation, only expression levels of *ELF1*, *GABPA*, *GABPB1*, and *GABPB2* were significantly positively correlated with *SDHD* mRNA levels in the subset of samples wild-type for the *SDHD* promoter (Fig. 8A and Supplementary Fig. S5). Among them, *ELF1* and *GABPA* were the two most significantly correlated transcription factors, with Pearson correlation coefficients of 0.46 ($P = 4.40 \times 10^{-8}$) and 0.42 ($P = 8.57 \times 10^{-7}$), respectively (Fig. 8B and 8C); there was a non-significant trend towards correlation between expression levels of *GABPA* and *SDHD* in samples harboring the C523T mutation (Pearson correlation coefficient 0.45, $P = 0.17$)

but not *ELF1* (Pearson correlation coefficient -0.03, P = 0.92). Significant correlations were not identified between the expression of *SDHD* and other transcription factors whose binding sites were predicted to be disrupted specifically by the C541T, C544T, as well as several other mutations in *SDHD* promoter wild-type TCGA SKCM samples (data not shown). In summary, these data are consistent with a potential role for GABPA, ELF1 and/or other ETS transcription factors in mediating *SDHD* expression.

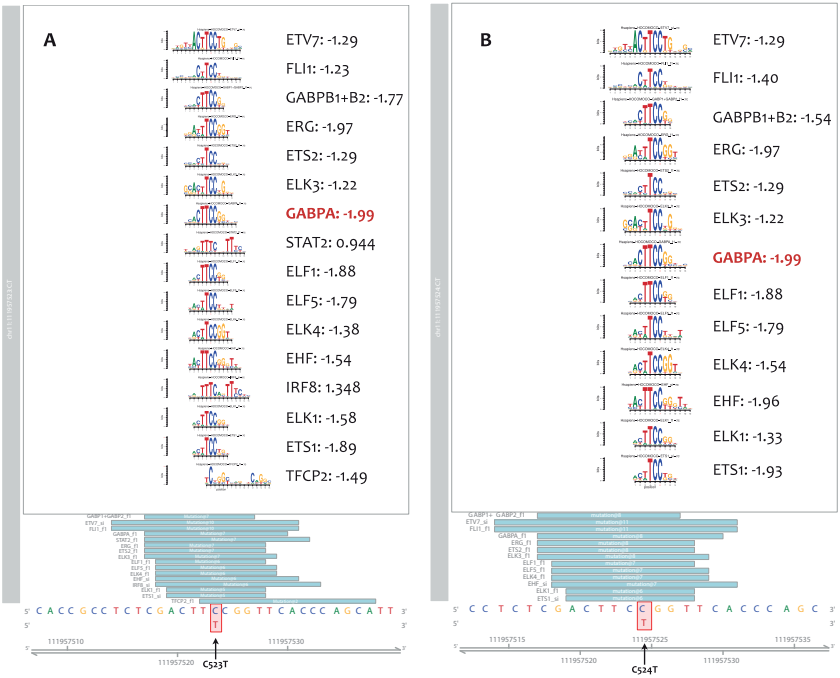


Figure 7. Predicting SDHD promoter mutation effects on transcription factor binding sites.

Data are shown for A) the C523T mutation and B) the C524T mutation. Genomic sequence and coordinates are at the bottom of the display; the positions of the matches represented (light blue boxes). The position of the mutation within the motif is indicated by a red-bounding box, with the alternate allele below in red font as on the motif logo position bar above. The motif logos generated from motifstack are shown above using the color conventions of the genomic sequence below. Predicted transcription factor name and change score (Alterscore-Refscore) are shown to the right of each motif, and the transcription factor with the strongest score is highlighted in red font. Mutations leading to predicted disruption of transcription factor binding have negative change scores, while those creating new transcription factor binding sites will have a positive change scores.

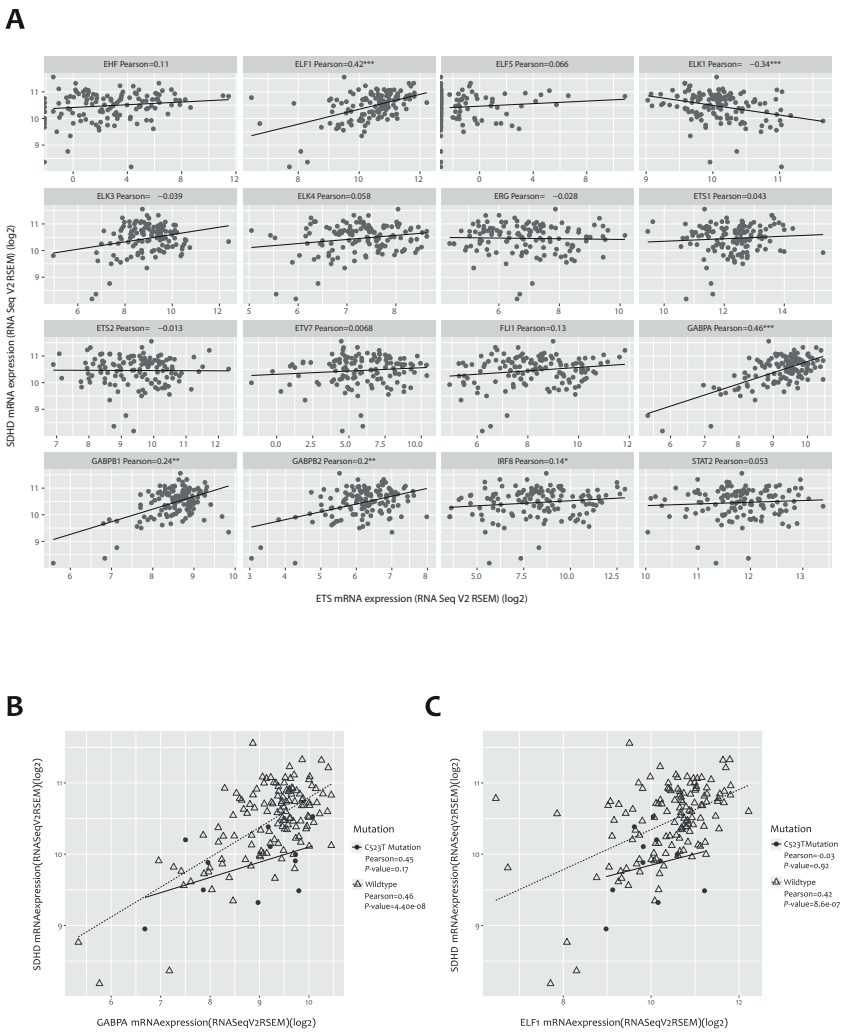


Figure 8. mRNA expression correlation between SDHD and multiple ETS transcription factors in SDHD promoter wild-type TCGA SKCM samples.

A Pearson correlation of mRNA expression between SDHD and 16 ETS transcription factors predicted by motifbreakR. Significant Pearson correlations are denoted with one or more star (*: P-value <0.05; **: P-value <0.01; ***: P-value <0.001).

B-C SDHD mRNA expression is highly correlated with B) GABPA and C) ELF1 mRNA expression specifically in SDHD promoter wild-type SKCM samples.

Identification of GABPA and GABPB1 as proteins preferentially binding the wild-type SDHD promoter by quantitative mass spectrometry

To perform an unbiased search for protein-DNA interactions specifically altered by *SDHD* promoter mutations, we used previously established workflows for AP-MS/MS based identification of sequence specific protein-DNA binding on a proteome-wide scale. Oligonucleotide baits were designed to encompass four mutation sites (C523T, C524T, C541T, and C544T) concurrently, and each oligo contained a mutation at one of these sites (Supplementary Table S3). All these mutations were located within the core motif of multiple ETS transcription factors as predicted by motifbreakR *in silico*. AP-MS/MS analysis of DNA pulldowns was performed using the metastatic melanoma-derived cell line UACC903. Interestingly, we identified components of the GABP transcription factor complex, GABPA and GABPB1, as wild-type specific interactors of the recurrent (C523T and C524T) *SDHD* promoter sites (Fig. 9A and 9B). GABP is unique amongst the ETS factors in that it alone is an obligate multimeric protein complex[58-60], where GABPA contains a DNA binding domain, yet the transcriptional activation domain is encoded by GABPB genes[58,61]. In addition, transcription factor ETS1 might also specifically interact with the wild-type sequence of both C523T and C524T sites, albeit to a lesser degree than GABPA and GABPB1. We did not identify any transcription factors specifically interacting with the wild-type or mutant sequence at mutation sites C541T and C544T (Fig. 9C and 9D).

To further confirm specific GABPA/GABPB1 binding at wild-type *SDHD* promoter mutation sites, we performed band-shift analysis of recombinant protein-DNA interactions using the same oligos previously used for AP-MS/MS analysis and recombinant human GABP or ELF1. Consistent with our results from AP-MS/MS analysis, we observed a specific and robust interaction of GABP (GABPA/B1 combined) at both the melanoma-specific C523T and C524T mutation sites via band-shift with recombinant protein (Fig. 9E and 9F). Intriguingly, band-shift experiments also revealed a present but relatively lower preference for the wild-type over the C541T and C544T mutated sequence (Fig. 9G and 9H). In addition, the band-shift experiments using recombinant ELF1 also suggested a slight preference for the wild-type sequence at C523T and C524T, especially at lower concentration (Supplementary Fig. S6). These data suggest that the transcription factors GABPA and GABPB1 specifically bind to the wild-type *SDHD* promoter sequence at the C523 and C524 sites, with this

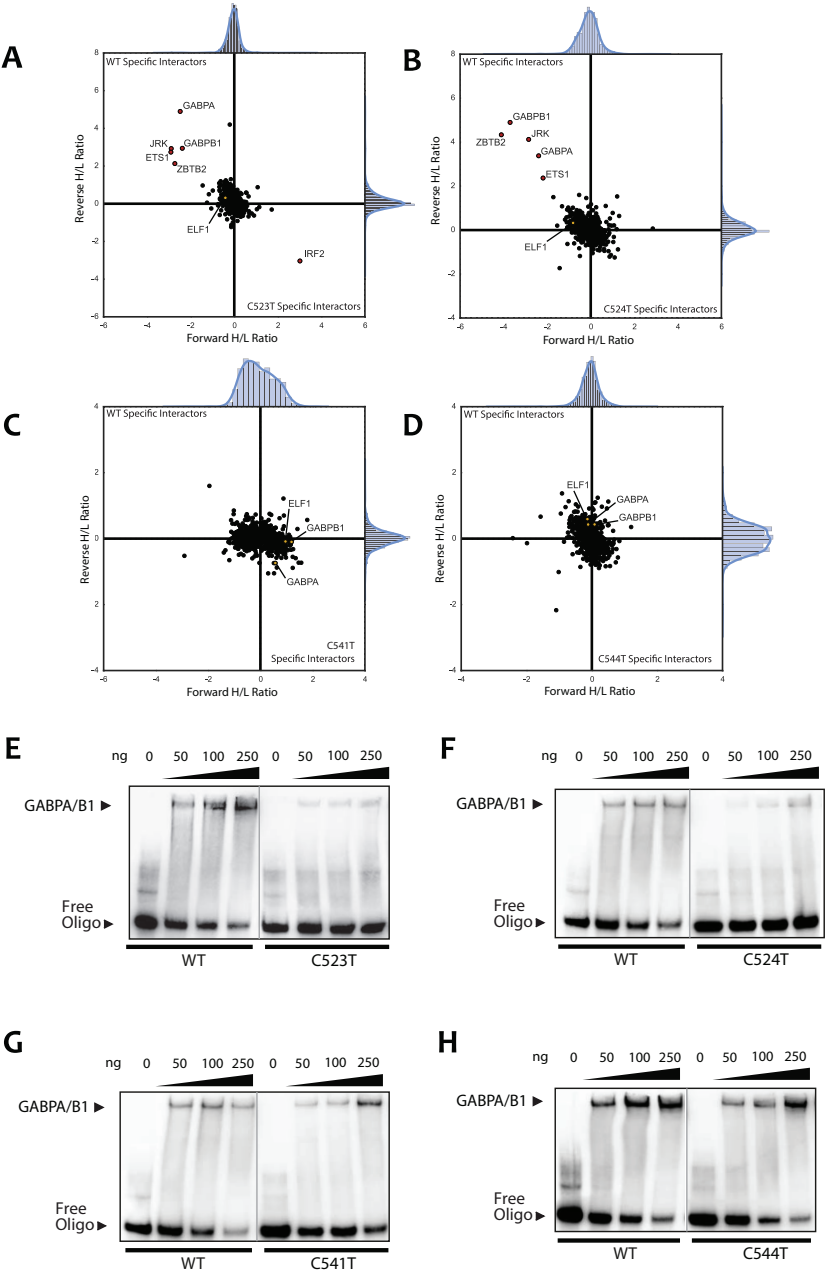

A
WT Specific Interactors
GABPA
JRK
GABPB1
ETS1
ZBTB2
ELF1
IRF2
C523T Specific Interactors
Reverse H/L Ratio
Forward H/L Ratio
B
WT Specific Interactors
GABPB1
ZBTB2
JRK
GABPA
ETS1
ELF1
C524T Specific Interactors
Reverse H/L Ratio
Forward H/L Ratio
C
WT Specific Interactors
ELF1
GABPB1
GABPA
C541T
Specific Interactors
Reverse H/L Ratio
Forward H/L Ratio
D
WT Specific Interactors
ELF1
GABPA
GABPB1
C544T Specific Interactors
Reverse H/L Ratio
Forward H/L Ratio
E
ng 0 50 100 250 0 50 100 250
GABPA/B1 ▶
Free Oligo ▶
WT
C523T
F
ng 0 50 100 250 0 50 100 250
GABPA/B1 ▶
Free Oligo ▶
WT
C524T
G
ng 0 50 100 250 0 50 100 250
GABPA/B1 ▶
Free Oligo ▶
WT
C541T
H
ng 0 50 100 250 0 50 100 250
GABPA/B1 ▶
Free Oligo ▶
WT
C544T

Figure 9. AP-MS/MS identifies allele-specific protein-DNA interactions for hotspot SDHD promoter mutations.

Custom oligos were 5'-biotinylated and designed to cover all analyzed mutation sites in the SDHD promoter as indicated in (Supplemental Table S3). Each mutation site was analyzed by two independent label-swapped experiments. ELF1, GABPA, and GABPB1, if not observed as significant interactors, are colored in yellow in the background cloud and noted by name.

A AP-MS/MS analysis of the C523T SDHD promoter mutation interactors. Interactors with a ratio of at least 3 IQR (inter-quartile range) in both replicate experiments are colored in red. Ratios are shown after log2 transformation. Labels were swapped between replicates to avoid labeling bias, hence specific interactors show a high ratio in one experiment and a low ratio in the other.

B AP-MS/MS analysis of the C524T SDHD promoter mutation interactors.

C AP-MS/MS analysis of the C541T SDHD promoter mutation interactors.

D AP-MS/MS analysis of the C544T SDHD promoter mutation interactors. E-H) band-shift experiments confirm SDHD WT-specific GABP binding at C523T and C524T mutation sites, but not C541T or C544T. Oligos were shifted using recombinant human protein as described in the materials and methods.

E band-shift analysis with the C523T SDHD promoter mutation oligo and recombinant human GABP (GABPA and GABPB1) protein. Band-shift experiments for each protein-oligo combination were resolved on the same gel at the same exposure. The grey line indicates where a single lane was cropped out for clarity.

F-H band-shift analysis with F) C524T G) C541T and H) C544T SDHD promoter mutation oligonucleotides and recombinant human GABP protein.

Regulation of SDHD expression by GABPA and GABPB1

To validate the potential regulation of *SDHD* expression by the ETS transcription factors GABPA, GABPB1, and ELF1, we knocked down the expression of these factors via siRNA in multiple melanoma cell lines (UACC903, UACC1113 and C021). Consistent with a role for GABPA in regulating *SDHD* expression, we observed that depletion of *GABPA* resulted in significantly lower mRNA expression of *SDHD* in all three of the cell lines tested five days following siRNA transfection, with average normalized expression of 0.50 relative to a scrambled siRNA control (range 0.37 - 0.71; Fig. 10A). In contrast, we did not detect a consistent effect of *GABPB1* depletion on levels of *SDHD* mRNA (Fig. 10B); while depletion of GABPB1 resulted in a subtle reduction of *SDHD* expression in UACC903, GABPB1 knockdown had the opposite effect in UACC1113. These data suggest that although GABPB1 was identified as a protein binding preferentially to the wild-type *SDHD* promoter, other proteins may compensate for the loss of GABPB1, and a GABP complex specifically composed of both GABPA and GABPB1 may not be the sole GABP complex regulating *SDHD*. Intriguingly, we found

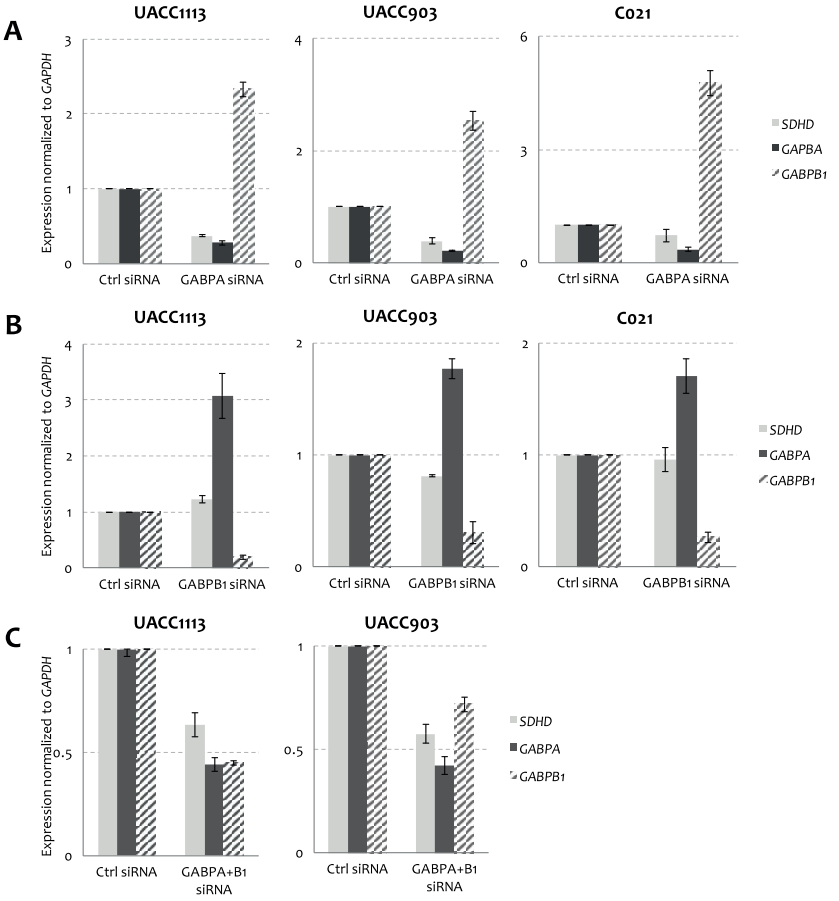


Figure 10. siRNA-mediated knockdown of GABPA and GABPB1 deregulates SDHD expression in melanoma cells.

A GABPA depletion decreases SDHD and increases GABPB1 expression in three melanoma cell lines (UACC1113, UACC903, and C021).

B GABPB1 depletion increases GABPA expression in three melanoma cell lines (UACC1113, UACC903 and C021), but has little effect on SDHD levels.

C concomitant depletion of both GABPA and GABPB1 decreases SDHD expression in both UACC1113 and UACC903 cell lines.

that depletion of either *GABPA* or *GABPB1* resulted in an increase in mRNA levels of the other (Fig. 10A and 10B), raising the possibility that any effect of *GABPB1* depletion on *SDHD* expression could have been masked by increased levels of GABPA. Further, concomitant depletion of both transcription factors

result in a consistent decrease of *SDHD* at both the mRNA (Fig. 10C). In contrast, depletion of *ELF1* had no effect when tested in multiple cell lines (Supplementary Fig. S7). Knockdown of *PRDM1* and *IRF4*, whose motifs are created by C541T and C544T mutations, respectively, resulted in varied but considerably lesser effects on *SDHD* expression across multiple cell lines than did depletion of *GABPA* (Supplementary Fig. S8A); luciferase reporter assays following knockdown of both genes suggested no change in expression for either mutation relative to that of wild-type (Supplementary Fig. S8B). Taken together, these data establish GABP as a major transcriptional regulator of *SDHD*.

Discussion

Understanding dynamic, specific interactions between transcription factors and the cognate DNA motifs they bind is crucial towards elucidating their mechanisms of transcriptional regulation. For TERT promoter mutations, understanding these dynamic binding specificities is of particular importance as transcriptional activation is directly correlated with oncogenic outcomes. This study utilizes an unbiased proteome-wide approach to identify dynamic interplay between GABP and ELF1 via their specific interactions with the TERT promoter mutations. This study also highlights the importance of considering both native motifs and the spatial architecture of DNA baits when performing DNA-protein AP-MS/MS experiments. The combinatorial nature of transcription factor binding is often difficult to predict, and this study offers a cautionary note that oligo baits of different lengths or different sequence/motif compositions might produce discrepant results even for the same locus. Indeed, as this study indicates, when using oligos designed to minimally cover a SNP or a cancer mutation, biologically important interactions may be overlooked. On the other hand, the importance of the super-structure (in terms of motif co-occurrence and spatial architecture) of promoter and enhancer elements is of growing interest in the study of gene regulatory mechanisms. This study shows that *in vitro*, MS-based approaches can shed light into combinatorial interactions between transcription factors that depend intrinsically not only on the motif interrupted or created by a variant but also on the presence of local (or distal) co-regulatory motifs.

Although most ETS-transcription factors are reported to bind monomerically, cooperative, antagonistic, and combinatorial effects are common[62]. For example, Bell et al. use an elegant bioinformatics analysis to indicate that strong GABPA binding sites from ENCODE ChIP-seq tracks correlates with multiple motifs spaced in a manner suggesting dimerization[29,63]. This periodic feature is unique to GABPA and was not seen in ELF1 ChIP-seq tracks, suggesting that ELF1 in contrast indeed binds monomerically. Indeed, GABP is known to be the only obligate multimer among ETS-family transcription factors[31,33,34]. Intriguingly, competitive binding between GABP and another ETS factor, PU.1, has been seen at the CD18 promoter, yet this competition cooperatively drives CD18 expression[56]. Recent reports have indicated that diverse combinatorial effects are crucial in the activating function of the TERT promoter mutations, specifically in cooperation with native ETS-motif binding factors[29,57]. Although our data indicates that ELF1 knockdown has only a mild effect on downstream TERT expression (which seems to be driven predominantly by GABP), we cannot exclude the possibility of combinatorial effects with ELF1 or other natively binding factors in the TERT core promoter.

This study highlights the dynamic nature of transcription factor binding at the recurrent TERT promoter mutations and points towards a complex picture of oncogenic transcriptional regulation at this locus. We present a model where stable, heterotetrameric, transcriptionally active GABP excludes monomeric ELF1 (Fig. 4). In doing so, we confirm the identification of GABP as a TERT mutation-specific interactor in melanoma and contribute to our knowledge of regulatory mechanisms at this important locus.

Searching for recurrent mutations occurring in a similar ETS-family signature context, Weinhold and colleagues reported recurrent SDHD promoter mutations in melanoma and reported a correlation between mRNA levels of the ETS transcription factor *ELF1* and *SDHD*, suggesting ELF1 as a potential key mediator of *SDHD* expression. By analyzing expression data for multiple ETS factors in the updated TCGA SKCM dataset, we observed a positive correlation between the levels of *SDHD* and multiple ETS transcription factors including *GABPA* and *ELF1*. Consistent with a potential role for GABPA in regulation of *SDHD*, motif analysis of the relatively common C523T and C524T mutations revealed that while these mutations indeed alter consensus sequences for numerous ETS factors, including ELF1, these mutations are predicted to most

strongly disrupt binding of GABPA. Taken together, these data suggest GABPA as a potential transcriptional regulator of the *SDHD* promoter.

We applied a mass spectrometry-based approach to identify proteins that preferentially bind to the wild-type *SDHD* promoter sequence as compared to several of the most commonly recurring promoter mutations (C523T, C524T, C541T, and C544T). Consistent with the motif analysis, GABPA was identified as a wild-type promoter interacting protein with binding disrupted specifically by the C523T and C524T mutations, in addition to GABPB1 and ETS1. In contrast, ELF1 did not show significant allele-preferential binding for either of these mutations. Nonetheless, both recombinant GABPA/B1 and ELF1 alone showed considerably decreased binding to oligos containing either of these two mutations. Analysis of the C541T and C544T mutations, which occur adjacent to or within a different "TTCC" motif, on the other hand, did not reveal statistically significant allele-specific binding proteins. Still, both mutations exhibited an allelic preference for both recombinant GABPA/B1 and ELF1, albeit more subtle than that observed for C523T/C524T. The differences observed between binding of recombinant proteins and those within crude lysates suggest that while both GABPA/B1 and ELF1 show an allelic preference for all four mutations, the situation is likely to be considerably more complex in melanoma cells. The observed differences may be attributable to cooperative and or competitive effects between ETS factors.

Consistent with a potential role for GABPA in regulation of *SDHD*, depletion of *GABPA* in melanoma cell lines resulted in decreased *SDHD* transcription. Depletion of *GABPB1*, on the other hand, did not consistently reduce *SDHD* levels, suggesting that another protein, likely GABPB2, may compensate for the loss of GABPB1. In contrast to GABPA, depletion of *ELF1* had no effect on *SDHD* levels. In all, these data point to GABPA, but not ELF1, as a key regulator of *SDHD*. Together with recent data supporting a role for GABPA in activating *TERT* in conjunction with recurrent promoter mutations in melanoma, these data raise the possibility that creation or alteration of GABPA binding motifs may be a more common mutational mechanism with functional consequences in melanoma.

References

1 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

2 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

3 Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502-506, doi:10.1038/nature11071 (2012).

4 Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).

5 Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251-263, doi:10.1016/j.cell.2012.06.024 (2012).

6 Nikolaev, S. I. *et al.* Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat Genet* **44**, 133-139, doi:10.1038/ng.1026 (2011).

7 Krauthammer, M. *et al.* Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat Genet* **47**, 996-1002, doi:10.1038/ng.3361 (2015).

8 Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet* **44**, 1006-1014, doi:10.1038/ng.2359 (2012).

9 Shain, A. H. *et al.* Exome sequencing of desmoplastic melanoma identifies recurrent NFKBIE promoter mutations and diverse activating mutations in the MAPK pathway. *Nat Genet* **47**, 1194-1199, doi:10.1038/ng.3382 (2015).

10 Wong, S. Q. *et al.* Whole exome sequencing identifies a recurrent RQCD1 P131L mutation in cutaneous melanoma. *Oncotarget* **6**, 1115-1127, doi:10.18632/oncotarget.2747 (2015).

11 Dutton-Regester, K. *et al.* A highly recurrent RPS27 5'UTR mutation in melanoma. *Oncotarget* **5**, 2912-2917, doi:10.18632/oncotarget.2048 (2014).

12 Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-961, doi:10.1126/science.1230062 (2013).

13 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-959, doi:10.1126/science.1229259 (2013).

14 Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet* **47**, 710-716, doi:10.1038/ng.3332 (2015).

15 Harland, M. *et al.* Germline TERT promoter mutations are rare in familial melanoma. *Fam Cancer* **15**, 139-144, doi:10.1007/s10689-015-9841-9 (2016).

16 Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**, 1258-1263, doi:10.1038/ng.3141 (2014).

17 Denisova, E. *et al.* Frequent DPH3 promoter mutations in skin cancers. *Oncotarget* **6**, 35922-35930, doi:10.18632/oncotarget.5771 (2015).

18 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-1165, doi:10.1038/ng.3101 (2014).

19 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

20 Low, K. C. & Tergaonkar, V. Telomerase: central regulator of all of the hallmarks of cancer. *Trends in biochemical sciences* **38**, 426-434 (2013).

21 Wright, W. E., Piatyszek, M. A., Rainey, W. E., Byrd, W. & Shay, J. W. Telomerase activity in human germline and embryonic tissues and cells. *Dev Genet* **18**, 173-179, doi:10.1002/(SICI)1520-6408(1996)18:2<173::AID-DVG10>3.0.CO;2-3 (1996).

22 Greenberg, R. A. *et al.* Telomerase reverse transcriptase gene is a direct target of c-Myc but is not functionally equivalent in cellular transformation. *Oncogene* **18**, 1219-1226 (1999).

23 Yin, L., Hubbard, A. K. & Giardina, C. NF-κB regulates transcription of the mouse telomerase catalytic subunit. *Journal of Biological Chemistry* **275**, 36671-36675 (2000).

24 Zhang, Y., Toh, L., Lau, P. & Wang, X. Human telomerase reverse transcriptase (hTERT) is a novel target of the Wnt/β-catenin pathway in human cancer. *Journal of Biological Chemistry* **287**, 32494-32511 (2012).

25 Chiba, K. *et al.* Cancer-associated TERT promoter mutations abrogate telomerase silencing. *eLife* **4**, e07918 (2015).

26 Naxerova, K. & Elledge, S. J. Taking the brakes off telomerase. *eLife* **4**, e09519 (2015).

27 Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-961 (2013).

28 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-959 (2013).

29 Bell, R. J. *et al.* The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science* **348**, 1036-1039 (2015).

30 LaMarco, K., Thompson, C. C., Byers, B. P., Walton, E. M. & McKnight, S. L. Identification of Ets-and notch-related subunits in GA binding protein. *Science* **253**, 789-792 (1991).

31 Oikawa, T. & Yamada, T. Molecular biology of the Ets family of transcription factors. *Gene* **303**, 11-34 (2003).

32 Sawada, J.-i., Goto, M., Sawa, C., Watanabe, H. & Handa, H. Transcriptional activation through the tetrameric complex formation of E4TF1 subunits. *The EMBO journal* **13**, 1396 (1994).

33 Thompson, C. C., Brown, T. A. & McKnight, S. L. Convergence of Ets-and notch-related structural motifs in a heteromeric DNA binding complex. *Science* **253**, 762-768 (1991).

34 Chinenov, Y., Henzl, M. & Martin, M. E. The alpha and beta subunits of the GA-binding protein form a stable heterodimer in solution. Revised model of heterotetrameric complex assembly. *J Biol Chem* **275**, 7749-7756 (2000).

35 Stern, J. L., Theodorescu, D., Vogelstein, B., Papadopoulos, N. & Cech, T. R. Mutation of the TERT promoter, switch to active chromatin, and monoallelic TERT expression in multiple cancers. *Genes Dev*, doi:10.1101/gad.269498.115 (2015).

36 Scholz, S. L. *et al.* Analysis of SDHD promoter mutations in various types of melanoma. *Oncotarget* **6**, 25868-25882, doi:10.18632/oncotarget.4665 (2015).

37 Hollenhorst, P. C., McIntosh, L. P. & Graves, B. J. Genomic and biochemical insights into the specificity of ETS transcription factors. *Annu Rev Biochem* **80**, 437-471, doi:10.1146/annurev.biochem.79.081507.103945 (2011).

38 Sharrocks, A. D. The ETS-domain transcription factor family. *Nat Rev Mol Cell Biol* **2**, 827-837, doi:10.1038/35099076 (2001).

39 Butter, F. *et al.* Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet* **8**, e1002982 (2012).

40 Spruijt, C. G., Baymaz, H. I. & Vermeulen, M. in *Gene Regulation* 137-157 (Springer, 2013).

41 Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2**, 1896-1906, doi:10.1038/nprot.2007.261 (2007).

42 Lau, H. T., Suh, H. W., Golkowski, M. & Ong, S. E. Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics. *J Proteome Res* **13**, 4164-4174, doi:10.1021/pr500630a (2014).

43 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

44 Cox, J. *et al.* A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc* **4**, 698-705, doi:10.1038/nprot.2009.36 (2009).

45 Hubner, N. C. *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J Cell Biol* **189**, 739-754, doi:10.1083/jcb.200911091 (2010).

46 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-404, doi:10.1158/2159-8290.CD-12-0095 (2012).

47 Stark, M. S. *et al.* Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. *Nat Genet* **44**, 165-169, doi:10.1038/ng.1041 (2011).

48 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

49 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-33, doi:10.1002/0471250953.bi1110s43 (2013).

50 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

51 Kanchi, K. L. *et al.* Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun* **5**, 3156, doi:10.1038/ncomms4156 (2014).

52 Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847-3849, doi:10.1093/bioinformatics/btv470 (2015).

53 Kulakovskiy, I. V. *et al.* HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* **41**, D195-202, doi:10.1093/nar/gks1089 (2013).

54 Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**, D142-147, doi:10.1093/nar/gkt997 (2014).

55 Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-94, doi:10.1093/nar/gkh012 (2004).

56 Rosmarin, A. G., Caprio, D. G., Kirsch, D. G., Handa, H. & Simkevich, C. P. GABP and PU.1 compete for binding, yet cooperate to increase CD18 (beta 2 leukocyte integrin) transcription. *J Biol Chem* **270**, 23627-23633 (1995).

57 Li, Y. *et al.* Non-canonical NF-kappaB signalling and ETS1/2 cooperatively drive C250T mutant TERT promoter activation. *Nat Cell Biol* **17**, 1327-1338, doi:10.1038/ncb3240 (2015).

58 Bell, R. J. *et al.* Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science* **348**, 1036-1039, doi:10.1126/science.aab0015 (2015).

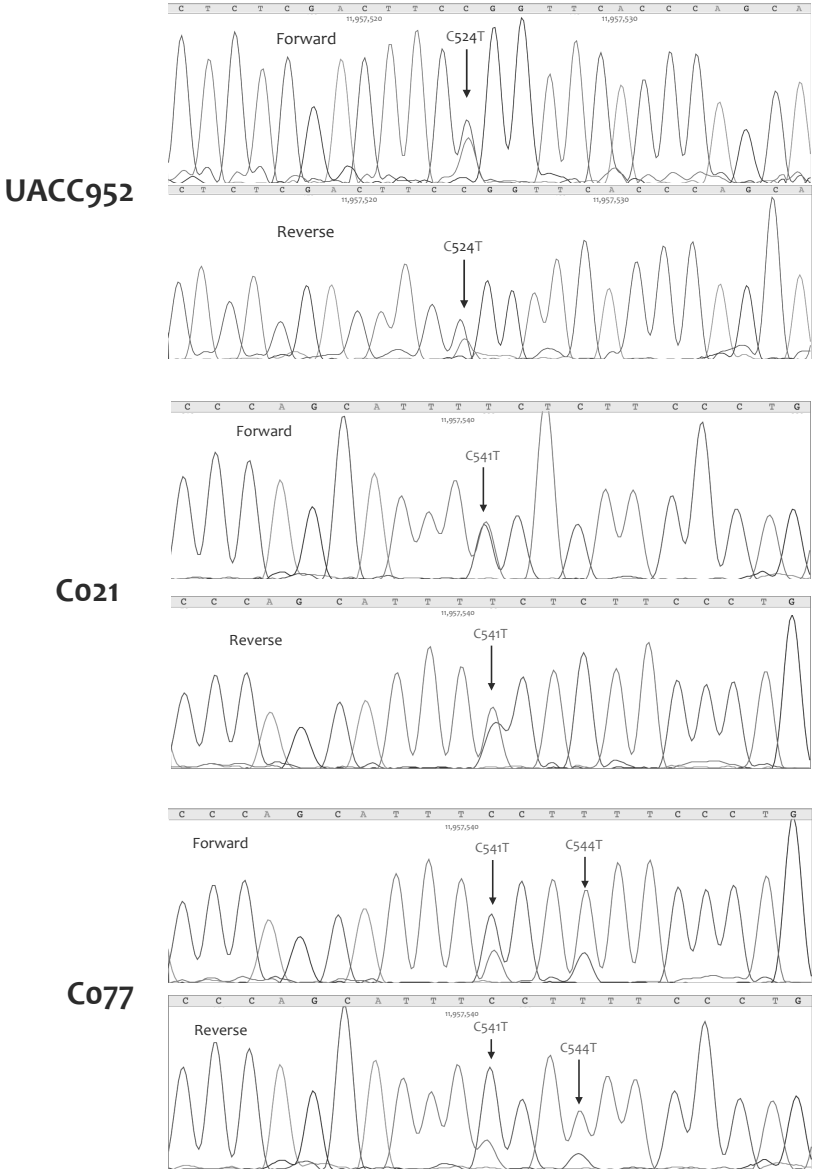
59 Yang, Z. F., Mott, S. & Rosmarin, A. G. The Ets transcription factor GABP is required for cell-cycle progression. *Nat Cell Biol* **9**, 339-346, doi:10.1038/ncb1548 (2007).

60 Thompson, C. C., Brown, T. A. & McKnight, S. L. Convergence of Ets- and notch-related structural motifs in a heteromeric DNA binding complex. *Science* **253**, 762-768 (1991).

61 LaMarco, K., Thompson, C. C., Byers, B. P., Walton, E. M. & McKnight, S. L. Identification of Ets- and notch-related subunits in GA binding protein. *Science* **253**, 789-792 (1991).

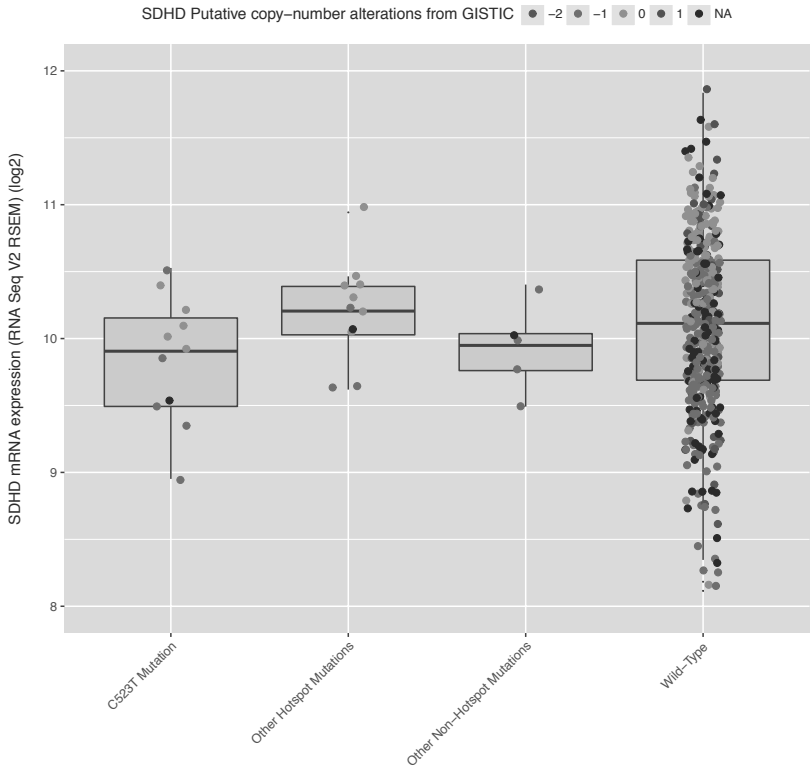
62 Li, R., Pei, H. & Watson, D. K. Regulation of Ets function by protein - protein interactions. *Oncogene* **19**, 6514-6523, doi:10.1038/sj.onc.1204035 (2000).

63 Consortium, E. P. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816, doi:10.1038/nature05874 (2007).

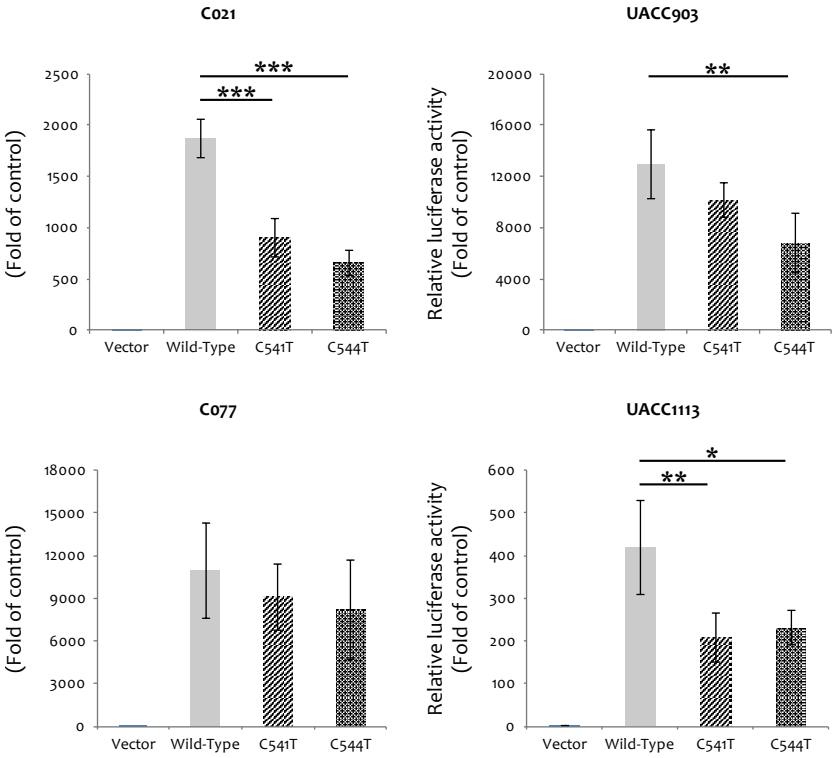


Supplementary Figure 1. Confirmation of the recurrent C524, C541 and C544 SDHD promoter mutations in melanoma cell lines.

Sanger sequencing traces for both forward and reverse reads verify all three recurrent SDHD promoter mutation in melanoma cell lines UACC952, C021 and C077.

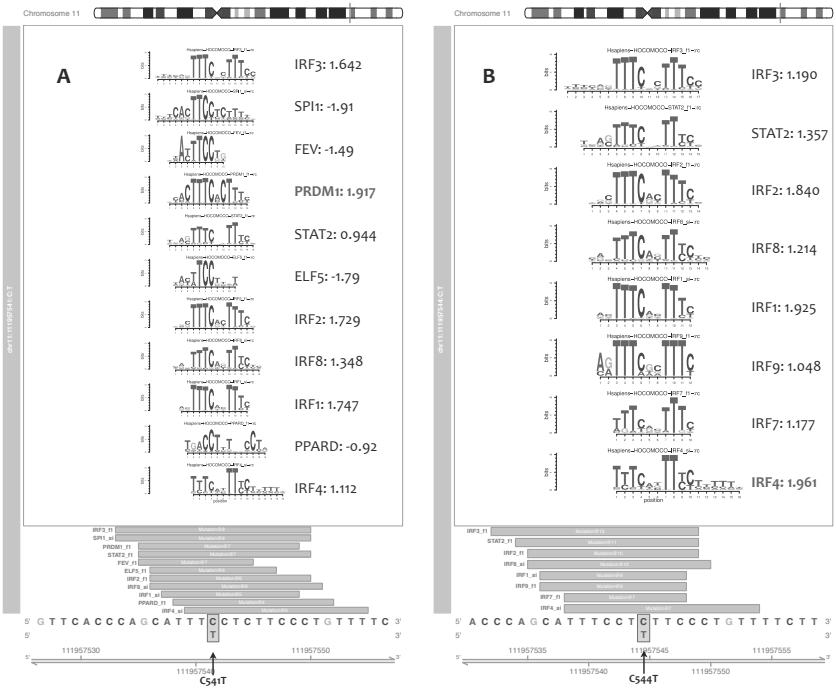


Supplementary Figure 2. SDHD expression difference in melanomas harboring promoter mutations compared to SDHD wild-type samples.
SDHD promoter-mutant samples are grouped according to mutation. SDHD expression was significantly decreased in the set of tumors harboring the SDHD C523T promoter mutation (onetailed student's t test, P = 0.0135) compared to wild-type samples. "Other Hotspot Mutation" includes the SDHD promoter mutations C532A, C541T, C544T and C548T, while "Non- Hotspot Mutation" includes SDHD promoter mutations C515T, A530G, C531T, C547T and C549T. Copy number for each sample as predicted by GISTIC is denoted with different coloring (Purple/-2: homozygous deletion; Red/-1: shallow deletion; Green/0: copy-neutral; Blue/1: copy gain; Black/NA: not assessable).



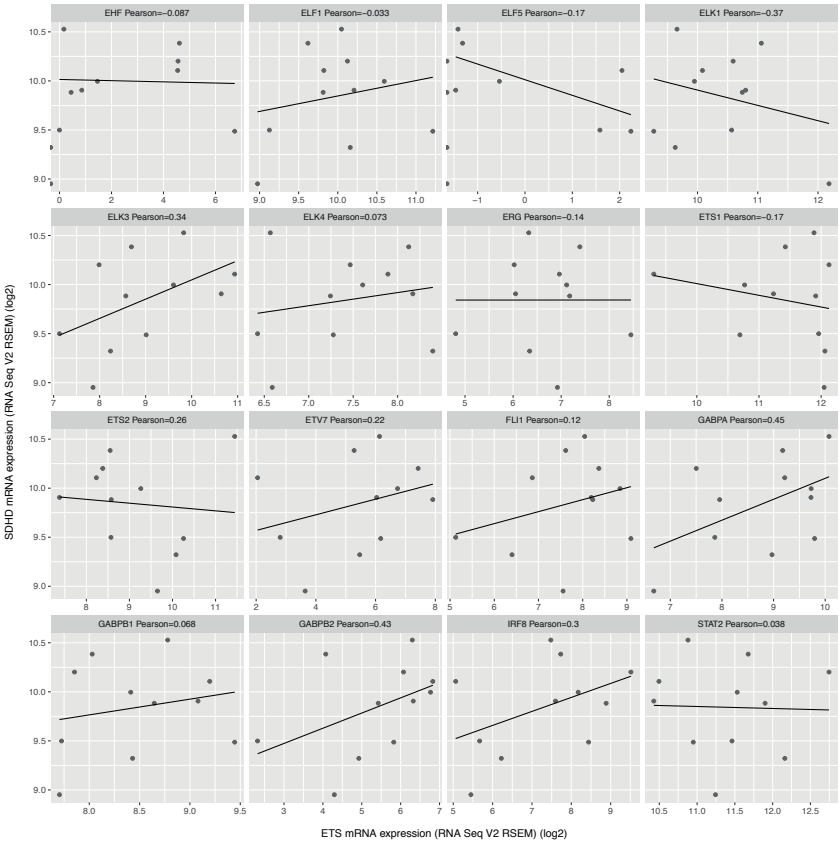
Supplementary Figure 3. SDHD promoter activity is significantly decreased by the C541T and C544T SDHD hotspot mutations in multiple melanoma cell lines.

A 163 bp fragment from the wild-type SDHD promoter sequence surrounding hotspot mutations significantly enhance luciferase reporter expression relative to vector control, whereas the same fragment containing hotspot mutations decrease enhancer activity relative to the wild-type sequence. Luciferase activity was measured 24hr after transfection and normalized to Renilla luciferase readings. Fold change over minimal promoter control (vector only) is plotted as relative luciferase activity. The experiment was performed four times with triplicates for each. Stars denote significant differences in luciferase activity by two-tailed student's t-test (*: P-value <0.05; **: P-value <0.01; ***: P-value <0.001).



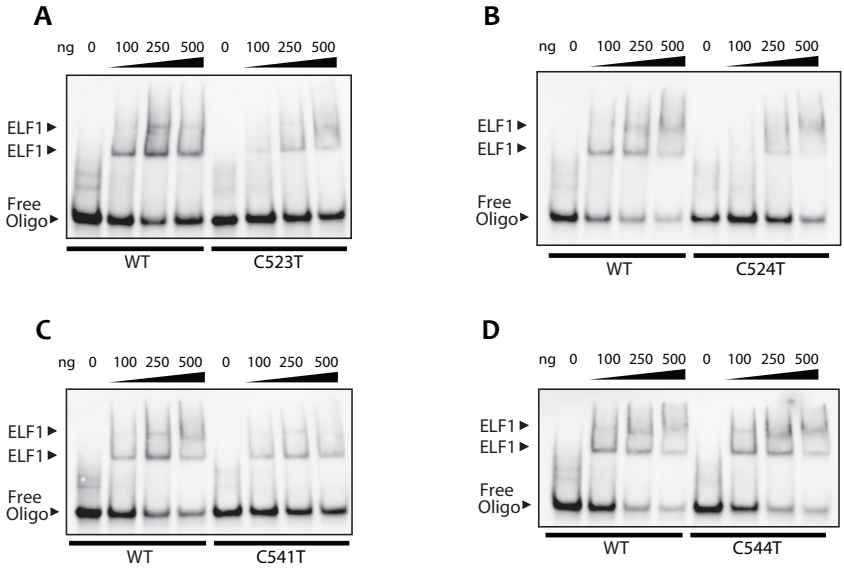
Supplementary Figure 4. Predicting SDHD promoter mutation effects on transcription factor binding sites.

Data are shown for (A) C541T and (B) C544T. Genomic sequence and coordinates are at the bottom of the display; the positions of the matches represented (light blue boxes). The position of the mutations within the motif is indicated by a red-bounded box, with the alternate allele below in red font as on the motif logo position bar above. The motif logos generated from motfstack are shown above using the color conventions of the genomic sequence below. Predicted transcription factor name and change score (Alterscore-Refscore) are shown to the right of each motif, and the transcription factor with the strongest score is highlighted in red. Mutations leading to disruption of transcription factor binding sites have negative change scores, while those creating new transcription factor binding sites have positive change scores.



Supplementary Figure 5. mRNA expression correlation between SDHD and multiple ETS transcription factors in TCGA SKCM samples harboring the SDHD C523T promoter mutation.

Pearson correlation of mRNA expression between SDHD and 16 transcription factors with consensus motifs altered by the C523T mutation as predicted by motifbreakR. Significant Pearson correlations are highlighted with one or more star (*: Pvalue <0.05; **: P-value <0.01; ***: P-value <0.001).



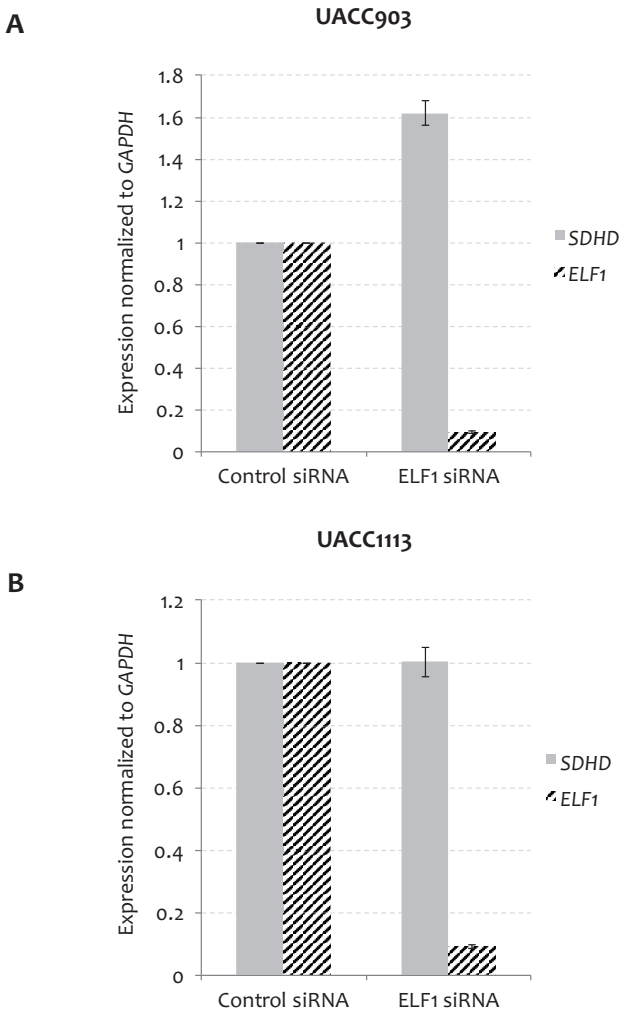
Supplementary Figure 6. Band-shift experiments indicate wild-type specific binding of ELF1 to the SDHD promoter.

A Band-shift analysis with the C523T SDHD promoter mutation oligo and recombinant human ELF1 protein.

B Band-shift analysis with the C524T SDHD promoter mutation oligo and recombinant human ELF1 protein.

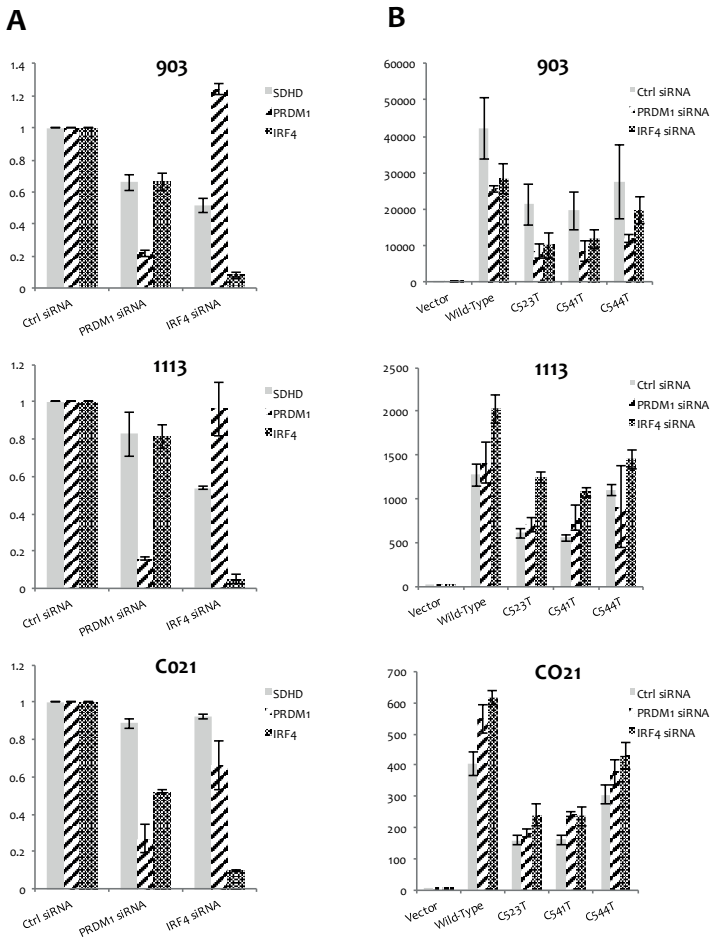
C Band-shift analysis with the C541T SDHD promoter mutation oligo and recombinant human ELF1 protein.

D Band-shift analysis with the C544T SDHD promoter mutation oligo and recombinant human ELF1 protein.



Supplementary Figure 7. siRNA-mediated knockdown of ELF1 does not lead to decreased SDHD expression in melanoma cells.

Control siRNA or siRNAs targeting ELF1 were transfected into cells, and expression of ELF1 and SDHD were assayed by Taqman assays at day 5 following transfection. Data are shown for two melanoma cell lines, A) UACC903 and B) UACC1113.



Supplementary Figure 8. Effects of siRNA-mediated knockdown of PRDM1 or IRF4 on SDHD expression and SDHD promoter activity in melanoma cells.

A Depletion of PRDM1 or IRF4 resulted in varied levels of reduction in SDHD expression across melanoma cell lines (UACC903, UACC1113, and C021). Control siRNA or siRNAs targeting PRDM1 or IRF4 were transfected into cells, and expression of PRDM1, IRF4 and SDHD were assayed by Taqman assays at day five following transfection.

B siRNA-mediated depletion of PRDM1 or IRF4 do not dramatically alter wild-type or mutant SDHD promoter activity in an allele-specific manner. A 163-bp fragment from the wild-type SDHD promoter sequence surrounding hotspot mutations significantly enhances luciferase reporter expression relative to vector control. The same fragment containing hotspot mutations results in decreased promoter activity relative to the wildtype sequence. While depletion of PRDM1 or IRF4 do broadly result in small alterations in reporter activity, neither alters reporter expression of these constructs in an allele specific manner. Fold change over minimal promoter control (vector only) is plotted as relative luciferase activity. The experiment was performed four times with triplicates for each.

Supplementary Table 1. SDHD promoter mutations identified in melanoma tumors datasets (TCGA, Broad and Yale) and melanoma cell lines.

Chromosome	Location	Ref	Alt	Sample	Source	Ref_bases_num	Alt_bases_num
11	111957518	G	A	Ma-Mel-114	Broad	24	16
11	111957527	T	C	Ma-Mel-35	Broad	12	3
11	111957529	C	T	JWCI-WGS-21	Broad	0	7
11	111957547	C	T	ME014	Broad	20	13
11	111957515	C	T	UACC257	Cell lines	257	12
11	111957517	C	T	C021	Cell lines	59	53
11	111957524	C	T	UACC1451	Cell lines	181	5
11	111957524	C	T	UACC2528	Cell lines	138	5
11	111957524	C	T	UACC952	Cell lines	78	83
11	111957541	C	T	C021	Cell lines	54	50
11	111957541	C	T	C077	Cell lines	35	11
11	111957544	C	T	C077	Cell lines	20	12
11	111957549	C	T	C025	Cell lines	8	19
11	111957556	C	T	C088	Cell lines	61	37
11	111957515	C	T	TCGA-EE-A29M	TCGA	8	10
11	111957523	C	T	TCGA-D3-A51G	TCGA	16	5
11	111957523	C	T	TCGA-D3-A8GI	TCGA	13	3
11	111957523	C	T	TCGA-D9-A1JW	TCGA	11	7
11	111957523	C	T	TCGA-DA-A1IC	TCGA	1	6
11	111957523	C	T	TCGA-EE-A29D	TCGA	6	3
11	111957523	C	T	TCGA-EE-A2GO	TCGA	13	8
11	111957523	C	T	TCGA-EE-A3J5	TCGA	8	3
11	111957523	C	T	TCGA-HR-A2OG	TCGA	5	6
11	111957523	C	T	TCGA-IH-A3EA	TCGA	4	4
11	111957523	C	T	TCGA-W3-AA1V	TCGA	7	4
11	111957523	C	T	TCGA-YD-A9TA	TCGA	3	6
11	111957530	A	G	TCGA-EB-A5UL	TCGA	18	3
11	111957531	C	T	TCGA-EE-A2GU	TCGA	8	5
11	111957532	C	A	TCGA-D3-A2JC	TCGA	12	3
11	111957532	C	A	TCGA-EE-A2A2	TCGA	165	3
11	111957532	C	A	TCGA-ER-A19B	TCGA	8	3
11	111957541	C	T	TCGA-EE-A2MD	TCGA	7	6
11	111957541	C	T	TCGA-EE-A2MI	TCGA	11	6
11	111957541	C	T	TCGA-FS-A1ZK	TCGA	0	8
11	111957544	C	T	TCGA-EE-A185	TCGA	6	7
11	111957544	C	T	TCGA-GN-A26C	TCGA	7	10
11	111957544	C	T	TCGA-W3-AA1Q	TCGA	12	8
11	111957547	C	T	TCGA-GN-A266	TCGA	12	9
11	111957548	C	T	TCGA-D3-A2JE	TCGA	21	8
11	111957548	C	T	TCGA-FR-A8YC	TCGA	2	9
11	111957549	C	T	TCGA-GN-A26C	TCGA	8	10
11	111957517	C	T	YUPADI	Yale	13	18
11	111957523	C	T	YUGEN8	Yale	0	43
11	111957523	C	T	YURIF	Yale	15	18
11	111957524	C	T	YUROO	Yale	16	23
11	111957535	G	T	YUFOLD	Yale	92	4
11	111957544	C	T	YUKLAB	Yale	60	29
11	111957544	C	T	YURUS	Yale	40	6
11	111957548	C	T	YUZEAL	Yale	14	28

Supplementary Table 2. MotifBreakR results predicting the effects of recurrent SDHD promoter mutations (C523T, C524T, C541T, C544T, and C548T) on transcription factor binding sites.

REF	ALT	snpPos	motifPos	geneSymbol	dataSource	providerName	providerId	seqMatch	pctRef	pctAlt	scoreRef	scoreAlt	Refpvalue	Altpvalue	alleleRef	alleleAlt	effect	dscore	dpct
C	T	111957544	10	IRF4	HOCOMOCO	IRF4_si	IRF4_HUMAN	tttcctCttcectgtt	0.749	0.931	8.314	10.276	2.406E-03	4.231E-06	0.001	0.998	strong	1.961	0.182
C	T	111957544	4	IRF1	HOCOMOCO	IRF1_si	IRF1_HUMAN	catttcctCttc	0.809	0.961	10.412	12.337	8.883E-04	2.271E-05	0.000	1.000	strong	1.925	0.152
C	T	111957541	8	PRDM1	HOCOMOCO	PRDM1_f1	PRDM1_HUMAN	gcatttCctcttcc	0.734	0.859	11.437	13.354	1.073E-03	4.125E-05	0.001	0.994	strong	1.917	0.125
C	T	111957544	5	IRF2	HOCOMOCO	IRF2_f1	IRF2_HUMAN	gcatttcctCttcc	0.790	0.934	10.294	12.135	8.060E-04	1.904E-05	0.000	0.995	strong	1.841	0.144
C	T	111957541	8	IRF1	HOCOMOCO	IRF1_si	IRF1_HUMAN	atttCctcttcc	0.782	0.919	10.065	11.812	1.455E-03	6.670E-05	0.017	0.983	strong	1.747	0.138
C	T	111957541	9	IRF2	HOCOMOCO	IRF2_f1	IRF2_HUMAN	catttCctcttcce	0.742	0.878	9.683	11.412	1.741E-03	7.600E-05	0.016	0.984	strong	1.730	0.135
C	T	111957541	9	IRF3	HOCOMOCO	IRF3_f1	IRF3_HUMAN	cagcatttCctcttccc	0.737	0.879	8.831	10.474	1.852E-03	6.460E-05	0.018	0.982	strong	1.642	0.142
C	T	111957544	5	STAT2	HOCOMOCO	STAT2_f1	STAT2_HUMAN	agcatttcctCttcc	0.760	0.907	7.258	8.616	1.963E-03	4.840E-05	0.031	0.928	strong	1.358	0.147
C	T	111957523	10	IRF8	HOCOMOCO	IRF8_si	IRF8_HUMAN	gacttCcggttcacc	0.748	0.885	7.629	8.977	2.225E-03	8.056E-05	0.055	0.945	strong	1.348	0.137
C	T	111957541	10	IRF8	HOCOMOCO	IRF8_si	IRF8_HUMAN	catttCctcttccct	0.805	0.943	8.194	9.542	6.155E-04	1.085E-05	0.055	0.945	strong	1.348	0.137
C	T	111957544	6	IRF8	HOCOMOCO	IRF8_si	IRF8_HUMAN	gcatttcctCttccc	0.809	0.933	8.232	9.447	5.572E-04	1.760E-05	0.030	0.922	strong	1.214	0.124
C	T	111957544	5	IRF3	HOCOMOCO	IRF3_f1	IRF3_HUMAN	ccagcatttcctCttcc	0.840	0.943	10.028	11.218	1.671E-04	7.514E-06	0.041	0.913	strong	1.190	0.103
C	T	111957544	4	IRF7	HOCOMOCO	IRF7_f1	IRF7_HUMAN	tttcctCttc	0.831	0.979	6.730	7.908	1.288E-03	1.240E-05	0.000	0.897	strong	1.178	0.149
C	T	111957541	14	IRF4	HOCOMOCO	IRF4_si	IRF4_HUMAN	ttCctcttccctgttt	0.831	0.934	9.198	10.310	2.983E-04	3.393E-06	0.069	0.890	strong	1.113	0.103
C	T	111957544	4	IRF9	HOCOMOCO	IRF9_f1	IRF9_HUMAN	catttcctCttc	0.771	0.893	7.091	8.140	6.152E-04	2.259E-05	0.000	1.000	strong	1.049	0.122
C	T	111957541	9	STAT2	HOCOMOCO	STAT2_f1	STAT2_HUMAN	gcatttCctcttccc	0.818	0.920	7.793	8.737	4.709E-04	2.864E-05	0.116	0.856	strong	0.944	0.102
C	T	111957523	9	STAT2	HOCOMOCO	STAT2_f1	STAT2_HUMAN	cgacttCcggttcac	0.788	0.890	7.512	8.456	1.041E-03	8.519E-05	0.116	0.856	strong	0.944	0.102
C	T	111957548	5	FOXM1	HOCOMOCO	FOXM1_f1	FOXM1_HUMAN	cttcCctgtttttc	0.852	0.920	4.940	5.312	4.527E-04	3.801E-05	0.000	0.514	weak	0.372	0.068
C	T	111957544	8	ETS2	HOCOMOCO	ETS2_f1	ETS2_HUMAN	cctCttcectg	0.980	0.928	6.361	6.037	2.861E-05	8.788E-04	0.676	0.252	weak	-0.324	-0.052
C	T	111957548	4	PPARD	HOCOMOCO	PPARD_f1	PPARD_HUMAN	tttcctcttcCctg	0.879	0.813	7.460	6.930	3.036E-05	2.651E-04	0.770	0.056	strong	-0.530	-0.065
C	T	111957544	6	SPI1	HOCOMOCO	SPI1_si	SPI1_HUMAN	cagcatttcctCttccc	0.889	0.836	11.611	10.937	5.897E-06	5.725E-05	0.787	0.108	weak	-0.673	-0.053
C	T	111957541	11	PPARD	HOCOMOCO	PPARD_f1	PPARD_HUMAN	tttCctcttccctg	0.879	0.765	7.460	6.540	3.036E-05	9.085E-04	0.882	0.062	strong	-0.920	-0.113
C	T	111957523	6	ELK3	HOCOMOCO	ELK3_f1	ELK3_HUMAN	cgacttCcggtt	0.928	0.754	6.882	5.655	5.287E-05	3.725E-03	1.000	0.000	strong	-1.227	-0.174
C	T	111957524	5	ELK3	HOCOMOCO	ELK3_f1	ELK3_HUMAN	cgactteCggtt	0.928	0.754	6.882	5.655	5.287E-05	3.725E-03	1.000	0.000	strong	-1.227	-0.174
C	T	111957523	8	FLI1	HOCOMOCO	FLI1_f1	FLI1_HUMAN	tctegacttCcggttca	0.968	0.785	6.753	5.520	1.795E-05	4.244E-03	0.924	0.000	strong	-1.233	-0.184
C	T	111957523	8	ETV7	HOCOMOCO	ETV7_si	ETV7_HUMAN	tctegacttCcggttca	0.849	0.722	9.100	7.808	8.974E-05	1.963E-03	1.000	0.000	strong	-1.293	-0.127
C	T	111957524	7	ETV7	HOCOMOCO	ETV7_si	ETV7_HUMAN	tctegactteCggttca	0.849	0.722	9.100	7.808	8.974E-05	1.963E-03	1.000	0.000	strong	-1.293	-0.127
C	T	111957523	5	ETS2	HOCOMOCO	ETS2_f1	ETS2_HUMAN	cgacttCcggt	0.970	0.764	6.301	5.007	8.345E-05	1.436E-02	0.933	0.067	strong	-1.294	-0.206
C	T	111957548	4	ETS2	HOCOMOCO	ETS2_f1	ETS2_HUMAN	cctcttcCctg	0.980	0.768	6.361	5.034	2.861E-05	1.374E-02	0.936	0.015	strong	-1.327	-0.212
C	T	111957524	4	ETS2	HOCOMOCO	ETS2_f1	ETS2_HUMAN	cgactteCggt	0.970	0.759	6.301	4.974	8.345E-05	1.510E-02	0.936	0.015	strong	-1.327	-0.212
C	T	111957524	4	ELK1	HOCOMOCO	ELK1_f1	ELK1_HUMAN	actteCggt	1.000	0.816	7.395	6.055	0.000E+00	3.193E-03	0.919	0.020	strong	-1.340	-0.184
C	T	111957523	6	ELK4	HOCOMOCO	ELK4_f1	ELK4_HUMAN	gacttCcggtt	0.999	0.868	10.671	9.283	1.192E-06	2.401E-04	0.945	0.045	strong	-1.388	-0.131
C	T	111957524	7	FLI1	HOCOMOCO	FLI1_f1	FLI1_HUMAN	tctegactteCggttca	0.968	0.760	6.753	5.353	1.795E-05	6.113E-03	0.962	0.038	strong	-1.400	-0.209
C	T	111957541	4	FEV	HOCOMOCO	FEV_f1	FEV_HUMAN	gcatttCctc	0.957	0.779	8.291	6.792	7.153E-05	3.216E-03	1.000	0.000	strong	-1.498	-0.178
C	T	111957523	15	TFCP2	HOCOMOCO	TFCP2_f1	TFCP2_HUMAN	tCcggttcacccagca	0.903	0.662	5.881	4.383	6.354E-05	2.268E-02	1.000	0.000	strong	-1.498	-0.242
C	T	111957524	5	ELK4	HOCOMOCO	ELK4_f1	ELK4_HUMAN	gactteCggtt	0.999	0.853	10.671	9.128	1.192E-06	3.133E-04	0.963	0.010	strong	-1.544	-0.146
C	T	111957524	3	GABPB1+GABPB2	HOCOMOCO	GABP1+GABP2_f1	GABP1+GABP2_HUMAN	cgactteCgg	0.987	0.807	8.641	7.096	1.049E-05	2.362E-03	0.973	0.027	strong	-1.544	-0.180
C	T	111957523	8	EHF	HOCOMOCO	EHF_si	EHF_HUMAN	gacttCcggttca	0.939	0.811	11.452	9.905	8.956E-06	7.288E-04	0.940	0.000	strong	-1.547	-0.128
C	T	111957523	5	ELK1	HOCOMOCO	ELK1_f1	ELK1_HUMAN	acttCcggt	1.000	0.783	7.395	5.811	0.000E+00	5.070E-03	0.959	0.013	strong	-1.584	-0.217
C	T	111957523	4	GABPB1+GABPB2	HOCOMOCO	GABP1+GABP2_f1	GABP1+GABP2_HUMAN	cgacttCcgg	0.987	0.780	8.641	6.865	1.049E-05	4.217E-03	1.000	0.000	strong	-1.776	-0.207
C	T	111957523	6	ELF5	HOCOMOCO	ELF5_f1	ELF5_HUMAN	gacttCcggtt	0.952	0.723	7.604	5.814	6.890E-05	1.560E-02	1.000	0.000	strong	-1.790	-0.229
C	T	111957524	5	ELF5	HOCOMOCO	ELF5_f1	ELF5_HUMAN	gactteCggtt	0.952	0.723	7.604	5.814	6.890E-05	1.560E-02	1.000	0.000	strong	-1.790	-0.229
C	T	111957541	6	ELF5	HOCOMOCO	ELF5_f1	ELF5_HUMAN	catttCctctt	0.970	0.742	7.745	5.955	2.480E-05	1.425E-02	1.000	0.000	strong	-1.790	-0.229
C	T	111957523	5	ELF1	HOCOMOCO	ELF1_f1	ELF1_HUMAN	gacttCcggt	0.997	0.788	9.083	7.203	6.676E-06	4.254E-03	1.000	0.000	strong	-1.881	-0.209
C	T	111957524	4	ELF1	HOCOMOCO	ELF1_f1	ELF1_HUMAN	gactteCggt	0.997	0.788	9.083	7.203	6.676E-06	4.254E-03	1.000	0.000	strong	-1.881	-0.209

Supplementary Table 2. Continued

REF	ALT	snpPos	motifPos	geneSymbol	dataSource	providerName	providerId	seqMatch	pctRef	pctAlt	scoreRef	scoreAlt	Refpvalue	Altpvalue	alleleRef	alleleAlt	effect	dscore	dpct
C	T	111957523	5	ETS1	HOCOMOCO	ETS1_si	ETS1_HUMAN	acttCcggt	0.985	0.743	7.821	5.925	1.526E-05	1.568E-02	0.996	0.000	strong	-1.896	-0.243
C	T	111957541	9	SPI1	HOCOMOCO	SPI1_si	SPI1_HUMAN	cagcatttCctcttccc	0.889	0.739	11.611	9.698	5.897E-06	1.082E-03	0.994	0.005	strong	-1.913	-0.150
C	T	111957524	4	ETS1	HOCOMOCO	ETS1_si	ETS1_HUMAN	actteCggt	0.985	0.738	7.821	5.887	1.526E-05	1.577E-02	1.000	0.000	strong	-1.934	-0.247
C	T	111957524	7	EHF	HOCOMOCO	EHF_si	EHF_HUMAN	gacteCggttca	0.939	0.776	11.452	9.487	8.956E-06	1.537E-03	1.000	0.000	strong	-1.966	-0.163
C	T	111957523	5	ERG	HOCOMOCO	ERG_fl	ERG_HUMAN	cgacttCcggt	0.983	0.811	11.372	9.392	1.907E-06	9.186E-04	1.000	0.000	strong	-1.980	-0.172
C	T	111957524	4	ERG	HOCOMOCO	ERG_fl	ERG_HUMAN	cgacteCggt	0.983	0.811	11.372	9.392	1.907E-06	9.186E-04	1.000	0.000	strong	-1.980	-0.172
C	T	111957523	7	GABPA	HOCOMOCO	GABPA_fl	GABPA_HUMAN	cgacttCcggttc	0.983	0.820	12.201	10.205	3.353E-06	7.427E-04	1.000	0.000	strong	-1.995	-0.162
C	T	111957524	6	GABPA	HOCOMOCO	GABPA_fl	GABPA_HUMAN	cgacteCggttc	0.983	0.820	12.201	10.205	3.353E-06	7.427E-04	1.000	0.000	strong	-1.995	-0.162

Note:
REF :the reference allele for the SNP
ALT :the alternate allele for the SNP
snpPos :the coordinates of the SNP
motifPos :the coordinates of the SNP within the TF binding motif
geneSymbol :the geneSymbol corresponding to the TF of the TF binding motif
dataSource :the source of the TF binding motif
providerName, providerId :the name and id provided by the source
seqMatch :the sequence on the 5' -> 3' direction of the "+" strand that corresponds to DNA at the position that the TF binding motif was found.
pctRef :The score as determined by the scoring method, when the sequence contains the reference SNP allele, normalized to a scale from 0 - 1. If filterp = FALSE, this is the value that is thresholded.
pctAlt :The score as determined by the scoring method, when the sequence contains the alternate SNP allele, normalized to a scale from 0 - 1. If filterp = FALSE, this is the value that is thresholded.
scoreRef :The score as determined by the scoring method, when the sequence contains the reference SNP allele
scoreAlt :The score as determined by the scoring method, when the sequence contains the alternate SNP allele
Refpvalue :p-value for the match for the pctRef score, initially set to NA. see calculatePvalue for more information
Altpvalue :p-value for the match for the pctAlt score, initially set to NA. see calculatePvalue for more information
alleleRef :The proportional frequency of the reference allele at position motifPos in the motif
alleleAlt :The proportional frequency of the alternate allele at position motifPos in the motif
effect :one of weak, strong, or neutral indicating the strength of the effect.
dscore: scoreAlt-ScoreRef
dpct: pctAlt-pctRef

Supplementary Table 3. Oligonucleotide design for quantitative mass spectrometry.

Oligo name	Forward	Reverse
WT_SDHD	5'-GTGCACCGCCTCTCGACTTCcGGTTCACCCAGCATTTcCTcTTCCCTGTTTTCTTTCGTCG-3'	5'- CGACGAAAGAAAACAGGGAAgAGgAAATGCTGGGTGAACCgGAAGTCGAGAGGCGGTGCAC-3'
C524T	5'-GTGCACCGCCTCTCGACTTCtGGTTCACCCAGCATTTcCTcTTCCCTGTTTTCTTTCGTCG-3'	5'-CGACGAAAGAAAACAGGGAAGAGgAAATGCTGGGTGAACCaGAAGTCGAGAAGGCGGTGCAC-3'
C541T	5'-GTGCACCGCCTCTCGACTTCCGGTTCACCCAGCATTTcCTcTTCCCTGTTTTCTTTCGTCG-3'	5'-CGACGAAAGAAAACAGGGAAGAGaAAATGCTGGGTGAACCGGAAGTCGAGAAGGCGGTGCAC-3'
C544T	5'-GTGCACCGCCTCTCGACTTCCGGTTCACCCAGCATTTCCTtTCCCTGTTTTCTTTCGTCG-3'	5'- CGACGAAAGAAAACAGGGAAaAGGAAATGCTGGGTGAACCGGAAGTCGAGAAGGCGGTGCAC-3'

Chapter 3

A common intronic variant of PARP1 confers melanoma risk via regulation by RECQL of an allelic DNA structure

Modified from:
A common intronic variant of PARP1 confers melanoma risk and mediates melanocyte growth via regulation of MITF.

Jiyeon Choi*, Mai Xu*, Matthew M Makowski, Tongwu Zhang, Matthew H. Law, Michael A. Kovacs, Anton Granzhan, Wendy J Kim, Hemang Parikh, Michael Gartside, Jeffrey M. Trent, Marie-Paule Teulade-Fichou, Mark M. Iles, Julia A. Newton-Bishop, D. Timothy Bishop, Stuart MacGregor, Nicholas K Hayward, Michiel Vermeulen, Kevin M. Brown

Nature Genetics. 2017.

Abstract

Prior genome-wide association studies have identified a melanoma-associated locus on chr1q42.1 that encompasses a ~100 kb region spanning the PARP1 gene. eQTL analysis in multiple cell types of melanocytic lineage consistently demonstrated that the 1q42.1 melanoma risk allele (rs3219090, G) is correlated with higher PARP1 levels. In silico fine-mapping and functional validation identified a common intronic indel, rs144361550 (-/GGGCCC, r2 =0.947 with rs3219090) as displaying allele-specific transcriptional activity. A proteomic screen identified RECQL as utilizing an unusual sequence-independent binding mode to interact with rs144361550 in an allele-preferential manner. This study thus highlights a new role for RECQL-mediated allele-specific transcription of PARP1 in melanomagenesis.

Introduction

To date, genome-wide association studies (GWAS) have identified twenty common, genome-wide significant melanoma susceptibility loci[1-9], most of which do not appear to be explained by protein-coding variants. A subset of these loci harbor known pigmentation genes that mediate melanoma-associated phenotypes such as eye, hair, and skin color. While several loci harbor genes implicated in cancer, evidence directly linking common risk variants within most of these loci to altered function of specific genes is lacking.

MacGregor and colleagues initially identified a melanoma risk locus tagged by rs3219090 on chromosome band 1q42.1 in an Australian case-control study at a near genome-wide level of significance ($P = 9.3 \times 10^{-8}$, OR = 0.87, protective allele A)[8]. The association has since been replicated by multiple other studies[3,10], including most recently by a meta-analysis of 12,874 melanoma cases (rs1858550, $P = 1.7 \times 10^{-13}$)[7]. Notably, the locus at 1q42.1 has also been associated with melanoma survival[11], where the melanoma risk allele correlates with increased survival, an association that has since been replicated[12]. The region of association spans from 226.52 Mb to 226.63 Mb (hg19) of chromosome 1, encompassing the entirety of the poly(ADP-ribose) (PAR) polymerase-1 (*PARP1*) (OMIM: 173870) gene, and fine-mapping suggests that the association is best explained by a single-SNP model[3].

While a number of other genes are located in the vicinity of the association peak, *PARP1* has the most well-established role in cancer. PARP1 is best known for its role as a DNA repair enzyme and genotoxic sensor that functions in base excision repair (BER), single-strand break repair, and double-strand break repair[13]. Once PARP1 binds to damaged DNA, its enzymatic function is activated, and it covalently attaches PAR polymers to acceptor proteins, including histones and PARP1 itself[14]. PARP1 amplifies DNA damage signals, modifies chromatin structures to accommodate DNA damage response proteins, and further recruits DNA repair proteins[13,15,16]. While PARP1 is not directly involved in repair of UV signature mutations via nucleotide excision repair, its role in the repair of DNA lesions induced by reactive oxygen species (ROS) is well-established[17]. ROS are generated by UVA exposure[18], are a byproduct of melanin production[19], and appear to play a role in oncogene-induced senescence (OIS)[20,21]. Aside from DNA repair, PARP1 functions in regulating gene expression by modifying chromatin structure, associating with promoters and

enhancers, and acting as a transcriptional co-regulator[22,23]. While many of these roles rely on PARP1 catalytic activity, some are also PARylation-independent, as in the transcriptional co-regulator function for NF-kB and B-MYB[24,25].

In this study, we functionally characterized the 1q41.2 melanoma risk locus, demonstrating a consistent correlation of the risk genotype with levels of *PARP1* gene expression in tissues of melanocytic origin, identifying a gene regulatory variant within the first intron of *PARP1*, and elucidating a role for PARP1 in melanocyte OIS via regulation of the melanocyte master regulatory transcription factor, *MITF*.

Materials and Methods

Early passage melanoma cell line eQTL analysis

Early passage melanoma cell lines were obtained from the University of Arizona Cancer Center (UACC), and eQTL analysis was performed by combining gene expression profiling and SNP genotyping data. The use of cell lines was approved by National Institutes of Health Office of Human Subject Research. Early passage melanoma cell lines were grown in the medium containing RPMI1640, 10% FBS, 20 mM HEPES, and penicillin/streptomycin until ~70% confluent. All cell lines were tested negative for mycoplasma contamination. For RNA isolation, cells were washed twice with cold PBS on ice and lysed with Trizol. Trizol was heated to 65ºC for 5 min to maximize melanin removal. Following heating, 1 mL chloroform was added per 5 mL of Trizol, vortexed, cooled on ice for 5 min, and centrifuged. The aqueous phase was removed, and equal volume of 70% EtOH was added dropwise while vortexing at low speed. Ethanol /supernatant mixtures were added to Qiagen RNeasy midi columns, with the flow-through reapplied once. Samples were then processed per manufacturer's protocol. RNA quantity and integrity were assessed using Bioanalyzer, which yielded RIN>7 for all samples. Total RNA were expression profiled on Affymetrix U133Plus2 expression microarrays, with labeling, hybridization, washing, and scanning performed according to manufacturer's protocol. Background correction and quantile normalization of gene expression data were performed using Robust Multi-array Average (RMA) algorithm with the default settings (Affymetrix). For genomic DNA isolation, Qiagen DNeasy Blood and Tissue kit was used. DNA quantity was

measured using NanoDrop and PicoGreen fluorescent assay. All samples were profiled using Applied Biosystems Identifiler STR panel prior to genotyping on Illumina OmniExpress arrays. After quality assessment of genotypes samples with >0.1 missing rate were excluded from the analysis. Loci with > 0.1 missing rate, MAF < 0.01, or Hardy-Weinberg Equilbrium *P*- value < 5E-5 were also excluded. The genomic region encompassing 6Mb around the GWAS lead SNP rs3219090 (which was directly genotyped on the array) was imputed using IMPUTE2.2.2[26] and 1KG phase1 v3 April 2012 (build 37) as a reference data. After assigning imputed genotypes for 2 samples that were missing direct genotype for rs3219090 (recoded as 0.333 probability of each genotype), 59 total samples were qualified for eQTL analysis with gene expression and genotype data available. Affymetrix U133Plus2 annotates 17 genes and 11 other transcripts in the 2 Mb region centering at rs3219090. Among these, probes for 12 genes and 2 other transcripts passed QC, including PARP1. eQTL analysis was then performed for these samples and gene/transcripts using SNPTEST v2.5 (see URL section) considering an additive model for genotypes.

Allele discrimination qPCR

cDNA from early passage melanoma cell lines heterozygous for both rs3219090 and coding surrogate SNP rs1805414 as well as of normal genomic copy were assayed using custom-designed Taqman genotyping probe sets that do not recognize genomic DNA. To act as a standard, the same amplicon was PCR-amplified for each allele from cDNA and subsequently cloned into the pCR2.1 TOPO vector (Invitrogen) and sequence verified. A standard curve was then generated using known amounts of cloned amplicons by plotting 11 different points of allelic ratio against VIC/FAM signal ratio. Allele discrimination qPCR was performed in triplicate, and allelic ratio was calculated from the average ratio of VIC/FAM signal using the standard curve. Departure from expected allelic ratio (major/minor allele) of 1 was assessed using two-tailed Wilcoxon signed rank test.

Nomination of candidate functional variants

All LD r^2 values used for candidate variant nomination were from 1KG phase 3 data. r^2 values based on both the EUR and CEU populations were

considered to extract the maximum r^2 of each variant with the lead SNPs, rs3219090 and rs1858550. Meta-analysis *P*-values were obtained from the previously published work of Law and colleagues[7]; all samples used in the meta-analysis were collected with informed consent and ethics committee approvals as previously described. DHS peaks for the primary human melanocyte culture "melano" (ENCODE/Duke) were obtained from ENCODE database[27] through UCSC Genome browser (see URL section). DNase-seq data for penis foreskin melanocyte primary cell cultures "skin 01" and "skin 02" (shown as Melanocyte_1 and Melanocyte_2 in Fig 2 and Supplementary Fig 2) were obtained from Roadmap database (03/09/2015) and DHS peaks were called using MACS[28](see URL section) using the default settings and FDR 1% cutoff. DHS peak intervals were overlaid with the genomic position of each candidate variant to determine whether each candidate localizes within a DHS peak. Experimental duplicates for skin 01 (DS18590, DS18601) and skin 02 (DS19662, DS18668), and analytical duplicates for melano were used for our analysis. To call a variant to be within DHS in one sample, DHS overlapping the variant in either of the duplicates were counted. DHS peaks from DNase-seq data for two melanoma cell lines Mel2183 (ENCODE/Duke) and RPMI-7951 (ENCODE/UW) were obtained from the ENCODE database. Peaks from FAIREseq data for 11 melanoma culture samples were obtained from GEO (accession number: GSE60666). Histone mark annotation was performed in the same way. Primary melanocyte histone marks were taken from subsets of three individuals through Roadmap database (skin 01, skin 02, and skin 03; shown as Melanocyte_1, Melanocyte_2, and Melanocyte_3 in Fig 2 and Supplementary Fig 2). Source of each track visualized on UCSC genome browser is as follows: Melanocyte track names used for Fig 2 (Track 1: UCSF-UBC-USC Penis Foreskin Melanocyte primary Cells Histone H3K27ac Donor skin03 Library A15584 EA Release 9, Track 2: UCSF-UBC-USC Penis Foreskin Melanocyte primary Cells Histone H3K4me1 Donor skin03 Library A15579 EA Release 8, Track 3: UCSF-UBC-USC Penis Foreskin Melanocyte primary Cells Histone H3K4me3 Donor skin03 Library A15580 EA Release 8, Track 4: Melano DNaseI HS Density Signal from ENCODE/Duke, Track 5: Penis Foreskin Melanocyte Primary Cells Donor skin 01 DNase Uniformly Signal from Roadmap, Track 6: UW Penis Foreskin Melanocyte Primary Cells DNase Hypersensitivity Donor skin02 Library DNase DS19662 EA Release

9.), Melanoma track names used for Supplementary Fig 2 (Track 1 through 11: H3K27ac ChIP-seq signal from F-seq, Track 12 through 22: FAIRE-seq signal from F-seq, Track 23: Mel2183 DNaseI HS Density Signal from ENCODE/ Duke, Track 24: RPMI-7951 DNaseI HS Raw Signal Rep 1 from ENCODE/ UW.).

Luciferase reporter assays

Luciferase constructs were generated to include the DHS region encompassing each SNP. Sequences encompassing each variant were amplified from genomic DNA of HapMap CEU panel samples using the primers listed in Supplementary Table 14, cloned into pCR2.1TOPO vector, and subsequently cloned into pGL4.23 vector using EcoRV and HindIII enzymes or directly into pGL4.23 vector using primers with HindIII and XhoI sequence overhangs. Sequence-verified pGL4.23 constructs were then co-transfected with pGL4.74 (Renilla luciferase) into a melanoma cell lines UACC2331, UACC457, or UACC1308 using Lipofectamine 2000 reagent (Thermo Fisher) or electroporation with Lonza Amaxa SE kit and DS-150 protocol on 4D-Nucleofector system. Electoporation of primary human melanocytes (HEMn-LP, Invitrogen) was performed using the Lonza Amaxa P2 kit and protocol CA-137 (Lonza). When luciferase assays were combined with *RECQL* over-expression, empty pCMV6-Entry vector or human RECQL cDNA clone (Origene, RC200427) were co-transfected with luciferase constructs into HEK293FT cells using Lipofectamine 2000 reagent. Cells were collected 24hr following transfection and luciferase activity was measured using Dual-Luciferase reporter system (Promega) on GLOMAX multi detection system (Promega). All cell lines and primary cultures used here and onward were tested negative for mycoplasma contamination.

EMSA and antibody supershift

Nuclear extracts were prepared from actively growing normal human melanocytes (HEMn-LP, Invitrogen or melanoma cell lines (UACC) using NE-PER nuclear and cytoplasmic extraction kit (Thermo Scientific). DNA oligos for each variant were synthesized with 5' biotin labeling, and HPLC-purified (Life Technologies; probe sequences are listed in Supplementary Table 14). Forward and reverse strands were then annealed to make double stranded DNA probes.

Probes were bound to 0.5-4µg nuclear extracts pre-incubated with 1µg poly d(I-C) in binding buffer containing 10mM Tris (pH 7.5), 50 mM KCl, 1 mM DTT, 10 mM MgCl2, with or without 5% glycerol at 4ºC for 30min. Unlabeled competitor oligos were added to the reaction mixture 5min prior to the addition of probes. Completed reactions were run on 5% or 4-20% native acrylamide gel and transferred blots were developed using LightShift Chemiluminescent EMSA kit (Thermo Scientific) and exposed on film. Supershift antibodies (RecQL, sc-25547, Santa Cruz) or rabbit normal IgG (sc-2027, Santa Cruz) were bound to nuclear extract prior to poly d(I-C) incubation at 4 ºC for 1hr. EMSAs with purified recombinant protein were performed using RECQL (TP300427, Origene), where purified recombinant proteins were used in place of nuclear extract and poly d(I-C). Additional antibodies and recombinant proteins for validations are as follows. Antibodies are from Santa Cruz unless otherwise specified: anti-NCL (sc-8031), anti-SRSF3 (sc-13510), anti-CIRP (sc-161012), anti-BLM (sc-7790), anti-hnRNPD (sc-22368), anti-RBM3 (sc-162080), anti-TOP3A (sc-11257), anti-RPA1 (sc-14696), anti-DHX36 (A300-525A, Bethyl), anti-RPA3 (ab167593, Abcam). Recombinant proteins are from Origene: NCL (TP319082), CIRP (TP301639), RPA1(TP302066), hnRNPD (TP300660), RBM3 (TP760298).

Mass-spectrometry

Quantitative AP-MS/MS following SNP DNA pulldown and in-solution dimethyl chemical labeling was performed based on procedures described previously[29,30]. Nuclear extract from the melanoma cell line UACC2331 was collected as described previously using the Dignam lysis protocol[31]. For DNA pulldowns, 500 pmol of annealed, forward strand 5'-biotinylated oligo probes were coupled to streptavidin sepharose beads (GE Healthcare). Insertion and deletion allele probe sequences are listed in Supplementary Table 14. Beads were incubated with 450 µg of nuclear extract for 90 minutes plus 10 µg of non-specific competitor DNA (either 10 µg of poly-dAdT or 5 µg of poly-dAdT plus 5 µg poly-dIdC). After washes, beads were resuspended in 100mM TEAB buffer, proteins were reduced with 5mM TCEP, alkylated with 10 mM MMTS, and digested overnight with 0.25 µg trypsin. Digested peptides were labelled using in-solution dimethyl chemical labelling as described previously[32]. All experiments were performed in duplicate, and labels were swapped between

replicate pairs to prevent labeling bias. Heavy and light labelled peptides were mixed and prepared using C18-StageTips. Peptides were loaded on a column packed with 1.8 µm Reprosil-Pur C18-AQ beads (gift from Dr. Maisch) and eluted using a 120 minute gradient from 7%-32% buffer B (80% acetonitrile, 0.1% formic acid) at a flow rate of 250nL/min. Peptides were sprayed directly onto a Thermo QExactive mass spectrometer. Data was collected in top10 data-dependent acquisition mode. Thermo RAW files were analyzed with MaxQuant (version 1.3.0.5) by searching against the Uniprot curated human proteome. Methionine oxidation and N-terminal acetylation were considered as variable modifications and cysteine-dithiomethane was set as a fixed modification. Protein ratios normalized by median ratio shifting as described previously[33] were used for outlier calling. An outlier cutoff of 1.5 IQRs (inter-quartile ranges) in two out of two biological replicates was used.

For TMT experiments, DNA pulldowns were performed as described first by Hubner et al[30]. NaCl was replaced with LiCl in protein binding buffer for LiCl samples, and PhenDC3 or NC-Bis was added to the noted samples at a concentration of 20uM. After washes, samples were eluted into a 20% methanol buffer for reduction, alkylation, and digestion. Peptides were labelled using the 10-plex tandem mass tag (TMT) system (Thermo)[34] and measured on a Thermo Tribrid Fusion mass spectrometer using a 240 minute chromatography gradient essentially as described above. Thermo Proteome Discoverer (version 2.1) was used for peptide identification and reporter ion quantification.

Chromatin immunoprecipitation of RECQL

UACC2331 melanoma cells or primary human melanocytes (HEMn-LP, Invitrogen) were fixed with 1% formaldehyde when 80-90% confluent, following the instructions of Active Motif ChIP-ITexpress kit or ChIP-IT high sensitivity kit. 7.5 x 10^6 cells were then sheared by sonication using a Bioruptor (Diagenode) at high setting for 15min, with 30 sec on and 30 sec off cycles. Sheared chromatin from 1 to 4 x 10^6 cells were used for each immunoprecipitation with antibodies against RECQL, H110 (sd-25547; Santa Cruz), and A300 (A300-450A; Bethyl), or normal rabbit IgGs (sc-2027; Santa Cruz) following the manufacturer's instructions. Purified pulled-down DNA was assayed by SYBR Green qPCR for enrichment of target sites using primers listed in Supplementary Table 14. A commercial primer set (71001, Active

Motif) recognizing a gene desert on chromosome 12 was used for a negative control (Neg).

Overexpression of RECQL in melanoma cell lines

RECQL was cloned from a cDNA construct (RC200427, purchased from Origene) into lentiviral pLU-TCMV-FMCS-pPURO vector (a generous gift from Dr. Meenhard's lab at Wistar) containing tetracycline-inducible promoter. Lentiviral vectors were co-transfected into 293 cells with packaging vectors psPAX2, pMD2-G, and pCAG4-RTR2. Virus was collected two days after transfection and concentrated by Vivaspin. Cells were incubated with virus for 24 hr, followed by puromycin (1-2 μg/ml) selection for 2-3 days. After drug selection cells were seeded and grown in the media containing varying amount of doxycycline (0,0.5, and 1 μg/ml). Cells were harvested after 48 hrs of doxycycline induction for RNA and protein isolation. cDNA was generated from total RNA and transcript levels were measured using Taqman qPCR. *RECQL* and *PARP1* transcript levels normalized to the levels of *B2M*, and PCR triplicates were averaged and considered as one data point. Western blotting of RECQL and GAPDH was performed using the following antibodies: RecQL1, sc-25547, Santa Cruz, GAPDH, sc-51907, Santa Cruz. For western blot analysis, total cell lysates were generated with RIPA buffer (Thermo Scientific, Pittsburgh, PA) and subjected to water bath sonication. Samples were resolved by 4-12% Bis-Tris ready gel (Invitrogen, Carlsbad, CA) electrophoresis.

Statistical analyses

All cell-based experiments were repeated at least three times with separate cell cultures, except for Fig3a (repeated twice), Fig4d-e (six technical replicates). When a representative set was shown, replicate experiments displayed similar patterns. For all the plots, individual data points are shown with median or mean, range (maximum and minimum), and 25 & 75 percentile (where applicable). Statistical method, number of data points, and number and type of replicates are indicated in each figure legend.

Results

The rs3219090 risk allele is correlated with high PARP1

We performed expression quantitative trait locus (eQTL) analysis in order to identify genes for which expression levels are correlated with 1q42.1 risk genotype in tissues of melanocytic lineage. Initially we evaluated the correlation of rs3219090 with expression of genes within +/-1Mb in 59 early-passage melanoma cell lines using expression microarray data. The results indicated that the rs3219090 risk allele is associated with higher levels of *PARP1* expression ($P = 1.4 \times 10^{-3}$, linear regression; Fig. 1a). Notably, *PARP1* is the only gene in the region that passed a Bonferroni-corrected *P*-value threshold (corrected for 14 genes, $P < 3.6 \times 10^{-3}$; Supplementary Table 1), and this eQTL subsequently validated via qPCR assay ($P = 0.031$, linear regression; Supplementary Fig. 1a). We then sought independent replication of *PARP1* and other nominally significant eQTL genes ($P < 0.05$) in publicly available RNA-sequencing datasets for melanoma-relevant tissues. When 409 melanoma tumors from The Cancer Genome Atlas (TCGA) project (dbGAP Accession: phs000178.v9.p8) were tested, the melanoma risk allele of rs3219090 was again significantly correlated with higher *PARP1* expression levels ($P = 3.9 \times 10^{-3}$, linear regression using copy number as a covariate; Supplementary Fig. 1b) while no other genes were significantly correlated (Supplementary Table 2). Similarly, the *PARP1* eQTL was replicated in normal skin samples collected through the Genotype-Tissue Expression (GTEx) Project (dbGAP Accession: phs000424.v6.p1), including those derived from both sun-exposed skin ($P = 2 \times 10^{-4}$, linear regression, n = 302) and non-sun-exposed skin ($P = 0.011$, linear regression, n = 196) (Supplementary Fig 1c-d, Supplementary Table 3-4). Together, these data identified *PARP1* as the strongest eQTL gene in the chr1q42.1 locus whose expression displayed the most consistent correlation with genotypes of the lead SNP in sample panels of melanocytic lineage as well as human skin.

To complement eQTL data and rule out the possibility of any sample-specific confounding factors masking genotype effect, we performed allele-specific expression (ASE) analysis for *PARP1* in samples carrying both risk and protective alleles. Fourteen melanoma cell lines that are heterozygous for rs3219090 and harbor normal regional copy number were assayed using

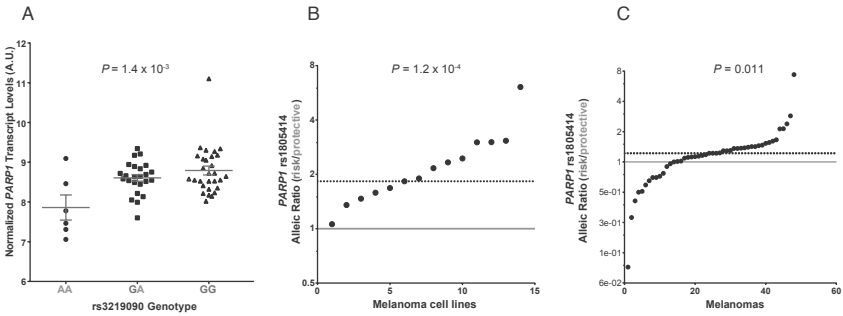


Figure 1. The melanoma risk-associated G allele of rs3219090 is correlated with increased *PARP1* expression.

A eQTL analysis was performed for rs3219090 using expression microarray and SNP array genotypes derived from a panel of 59 early-passage melanoma cell lines. A significant eQTL was observed for *PARP1*, and the result is plotted for rs3219090 genotype ($P = 1.4 \times 10^{-3}$; linear regression). G is the risk allele and A the protective allele of rs3219090. A.U.; arbitrary unit.

B The allelic ratios of *PARP1* transcripts were measured in 14 copy-neutral melanoma cell lines that were heterozygous for both rs3219090 and a synonymous mRNA-coding surrogate SNP (rs1805414, r^2=0.98 with rs3219090) using Taqman genotyping assays. Allelic ratios were inferred from a known amount of allelic standards and plotted as a ratio of *PARP1* expression from the risk over protective allele ($P = 1.2 \times 10^{-4}$, two-tailed Wilcoxon signed rank test, average value of PCR triplicates were considered as a single data point).

C Allelic ratios of *PARP1* transcripts were measured using RNA sequencing data from 48 copy-neutral TCGA skin melanoma samples that were heterozygous for both rs3219090 and rs1805414. The mapped numbers of RNAseq reads encompassing each allele of rs1805414 were used for calculating allelic ratios ($P = 0.011$, two-tailed Wilcoxon signed rank test). Solid line marks 1:1 ratio, and dashed line represents median ratio.

a quantitative allelic TaqMan assay for a synonymous coding surrogate SNP (rs1805414; $r^2 = 0.98$ with rs3219090 in 1KG phase3 EUR), where allelic ratio was inferred from known ratios of allelic standards. The results demonstrated a significant allelic imbalance towards a higher proportion of *PARP1* expressed from the risk allele in the majority of heterozygous cell lines ($P = 1.2 \times 10^{-4}$, two-tailed Wilcoxon signed rank test; Fig. 1b). Significant allelic imbalance was also observed when a subset of these cell lines were analyzed by RNAseq (data not shown). Subsequent *PARP1* ASE analysis in TCGA and GTEx RNAseq datasets demonstrated that a higher allelic proportion of mapped reads was also observed for the risk allele across TCGA tumor samples ($P = 0.011$, two-tailed Wilcoxon signed rank test, n = 48, copy-neutral and heterozygous; Fig. 1c), as well as in sun-exposed and non-sun-exposed skin tissues (GTEx, $P = 1.16 \times 10^{-5}$, n = 139; $P = 8.9 \times 10^{-5}$, n = 69; respectively, two-tailed Wilcoxon signed

rank test; Supplementary Fig. 1e-f). These data demonstrate that the melanoma risk allele of rs3219090 is significantly associated with increased *PARP1* expression in tissues of melanocytic origin and skin with striking consistency across multiple datasets.

Fine-mapping and functional annotation of candidate SNPs

Given that high *PARP1* levels are correlated with the melanoma risk allele of rs3219090, we next sought to identify functional risk variant(s) that may influence *PARP1* expression. Previously, fine-mapping of this locus in a large European population provided support for a model in which a single variant accounts for the association signal in this region[3], a finding confirmed as part of the meta-analysis conducted by Law and colleagues[7]. We prioritized 65 variants that are highly correlated with the lead SNP as candidate functional variants (r^2>0.6 with lead SNPs from the discovery or meta-analysis lead SNPs[3,7], rs3219090 and rs1858550, respectively; LD based on 1KG phase3, EUR and CEU). Given the absence of amino acid-changing *PARP1* variants within this set of candidates, an absence of evidence for alternative splicing as a likely mechanism (Supplementary Note), and considerable evidence for allelic differences in *PARP1* expression levels, we focused on those located within annotated melanocyte- or melanoma-specific *cis*-regulatory elements using data from the ENCODE[27] and Roadmap projects[35] (Supplementary Note, Supplementary Table 5-6, Supplementary Fig. 2-3). The four of the most strongly supported variants are situated at the center of melanocyte DHS peaks as well as within regions harboring promoter or enhancer histone marks (H3K4me1, H3K4me3, or H3K27ac) in the majority of melanocyte/melanoma cultures assayed (Supplementary Table 6). Based on these data, we proceeded with functional characterization of these four candidates (Supplementary Table 6, Supplementary Fig. 2).

An intronic indel displays allelic transcriptional activity

We assessed all four candidate functional variants for gene regulatory potential using luciferase reporter assays, as well as for allelic patterns of protein binding via electrophoretic mobility shift assay (EMSA). For these assays, we sought to identify variants that display 1) transcriptional activation consistent with ENCODE annotation, 2) higher activity for the risk allele consistent with

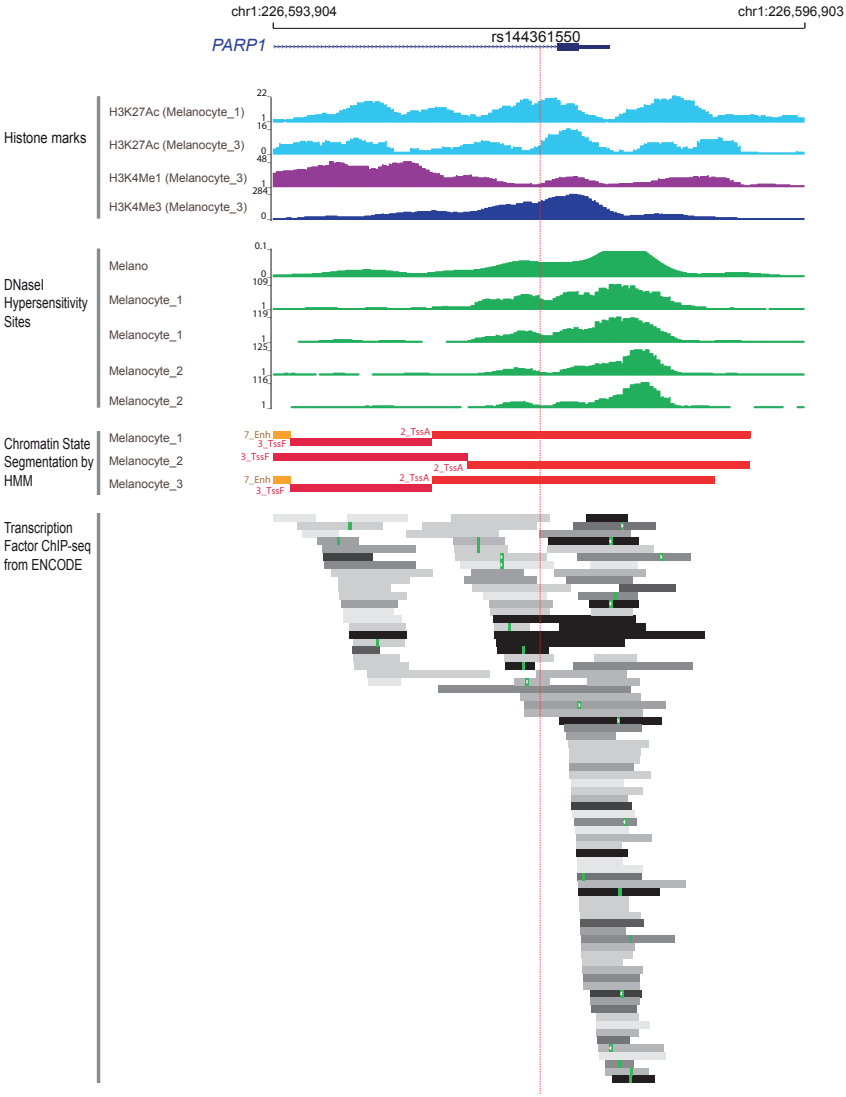


Figure 2. Functional annotation of a 3kb region encompassing rs144361550 in primary melanocytes.

Histone modifications (H3K4Me1, H3K4Me3, and H3K27Ac) and DNaseI hypersensitivity sites (DHS) in primary melanocytes are shown for a 3kb region encompassing rs144361550. The red dashed vertical line indicates the position of rs144361550, overlapping histone marks, DHS, and transcription factor binding sites. Genomic positions are based on hg19. Transcription factor binding sites are from UCSC genome browser track "Transcription Factor ChIP-seq (161 factors)

from ENCODE with Factorbook Motifs" representing multiple ENCODE cell types. "Chromatin Primary Core Marks Segmentation by HMM from Roadmap Project" track is also shown for three melanocyte samples. TssA: Active_TSS, TssF: Flanking_Active_TSS, Enh: Enhancers. For DHS, traces from two experimental replicates of Melanocyte 1 and 2 are displayed. The scale of each track is uniformly set throughout the region of the *PARP1* gene to cover the highest peaks, with 0 as the baseline (see online methods for details of each track).

the eQTL data, and 3) allele-specific protein binding. Among four candidate variants, only rs144361550, a GGGCCC indel variant, met all these criteria (Fig. 2-3, Supplementary Figs. 4-6; summarized in Supplementary Table 7). Namely, luciferase assays conducted in a melanoma cell line demonstrated that the genomic region around rs144361550 exhibits strong transcriptional activity in both long (905bp covering the larger DHS region, ~17-20 fold higher than control levels) and short cloned fragments (22 or 28bp covering the GGGCCC repeats, ~1.7-2.5 fold higher than control levels; Fig. 3a), where the risk-associated deletion allele exhibited higher reporter activity than the insertion allele (30-45% higher). In primary melanocytes, where transfection efficiency is considerably lower, allelic activity was not observed, but the long deletion and insertion fragments displayed weak but significant transcriptional activity (P = 1.2 x 10^{-3} and 5.9 x10^{-4}, respectively, two-tailed, paired t-test; Fig. 3c). EMSAs using nuclear extract from melanoma cell lines or cultured primary human melanocytes displayed preferential binding of nuclear proteins to the insertion allele (Fig. 3b,d). Given the potential for miscalling genotype of this functional indel, we directly genotyped rs144361550 in a large reference set to confirm LD with the lead SNP (Supplementary Note, Supplementary Table 8-9, Supplementary Fig. 7-8).

To identify proteins that bind rs144361550 in an allele-preferential manner, we utilized quantitative mass-spectrometry employing dimethyl label swapping[30,32 29]. Mass-spectrometry using melanoma cell line extract identified exclusively insertion allele-preferential interactors, the majority of which are not conventional transcription factors, including the RECQL helicase (Fig. 4a). While two transcription factors previously found by the ENCODE Project to localize to the region overlapping rs144361550 via chromatin immunoprecipitation (ChIP) were found to bind rs144361550 probes (TFAP2A, ZBTB7A), neither did so in an allele-preferential manner (data not shown), in line with the observation that rs144361550 creates no new sequence motifs but rather extends a poly-G repeat stretch. We then performed a series of antibody

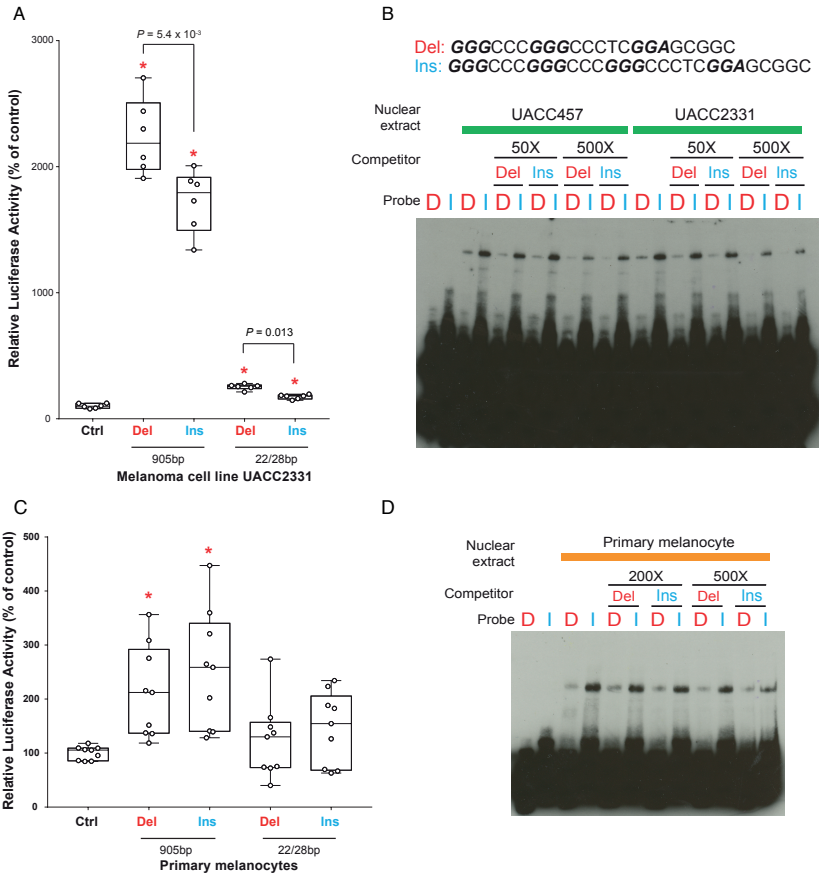


Figure 3. The melanoma-associated indel, rs144361550, drives allelic transcriptional activity and protein binding.

A,C Luciferase assays for rs144361550 were conducted using the melanoma cell line UACC2331 (A) or primary melanocytes (C). 905bp or 22/28bp encompassing rs144361550, respectively, were cloned 5' of minimal promoter in pGL4.23 vector and transfected into the cells. Luciferase activity was measured 24hrs after transfection and was normalized against Renilla luciferase activity. Relative luciferase levels were plotted as percent of the minimal promoter control. *P*-values are shown or * $P < 0.05$ against Ctrl (two-tailed, paired *t*-test). Two (A) and three (C) independent cell cultures and transfections of $n = 3$ were combined to present the median with range, 75 & 25 percentiles, and each data point. Ctrl: minimal promoter control, Del: deletion/risk allele construct, Ins: insertion/protective allele construct.

B, D EMSA was performed using biotin-labeled double stranded oligos for the deletion/risk (Del, D) or insertion/protective (Ins, I) alleles of rs144361550 and nuclear extracts from melanoma cell lines UACC2331 and UACC457 (B) or primary melanocytes (D). Sequences shown in panel b for the 22 bp deletion and 28 bp insertion probes were used for both b and d (bold and Italic bases highlight potential G4 structure forming nucleotides). 50X, 200X or 500X molar excess of unlabeled competitors were added in specified lanes.

supershifts and EMSAs using purified recombinant proteins for multiple candidates and validated that RECQL is an unequivocal allele-preferential binder to rs144361550 (Fig. 4b, Supplementary Figs. 9, and Supplementary Table 10). ChIP assays indicated that RECQL indeed binds to the *PARP1* indel region in melanoma cells and primary human melanocytes carrying an insertion allele (Fig. 4c, Supplementary Fig. 10). We also performed a series of *in silico* and *in vivo* assays testing for alternative DNA secondary structure formation (G-quadruplex or G4), with the results suggesting RECQL-specific allelic binding mechanism rather than the one through G4 (Supplementary Note, Supplementary Table 10-11, Supplementary Fig. 11-13). Intriguingly, chemical perturbation of potential ssDNA secondary structures either repelled proteins involved in (nucleotide excision repair) DNA damage associated processes or recruited proteins involved in RNA processing (Supplementary Fig. 14). These global effects were largely independent of allele, indicating a general structural mechanism. Yet, while RECQL prefers the insertion allele in a dsDNA context, RECQL prefers the deletion allele in a ssDNA, G4 permissive context, suggesting a complex and so-far uncharacterized DNA secondary structure.

Ectopic expression of RECQL in three melanoma cell lines carrying insertion or deletion alleles at a moderate level using lentiviral transduction resulted in a mild increase in *PARP1* transcription (Fig. 4d). We then performed luciferase assays for rs144361550 with or without RECQL over-expression in cells with low baseline levels of RECQL relative to melanomas (HEK293FT cells). At a basal level, insertion and deletion alleles did not display differential luciferase activity, but upon RECQL over-expression, significant allele-specific transcriptional activity we previously observed in melanoma cell lines was recapitulated (Fig. 4e). Together, these data suggest that RECQL may play a role in *PARP1* allelic expression in cells of melanocytic lineage through the melanoma risk-associated indel, rs144361550.

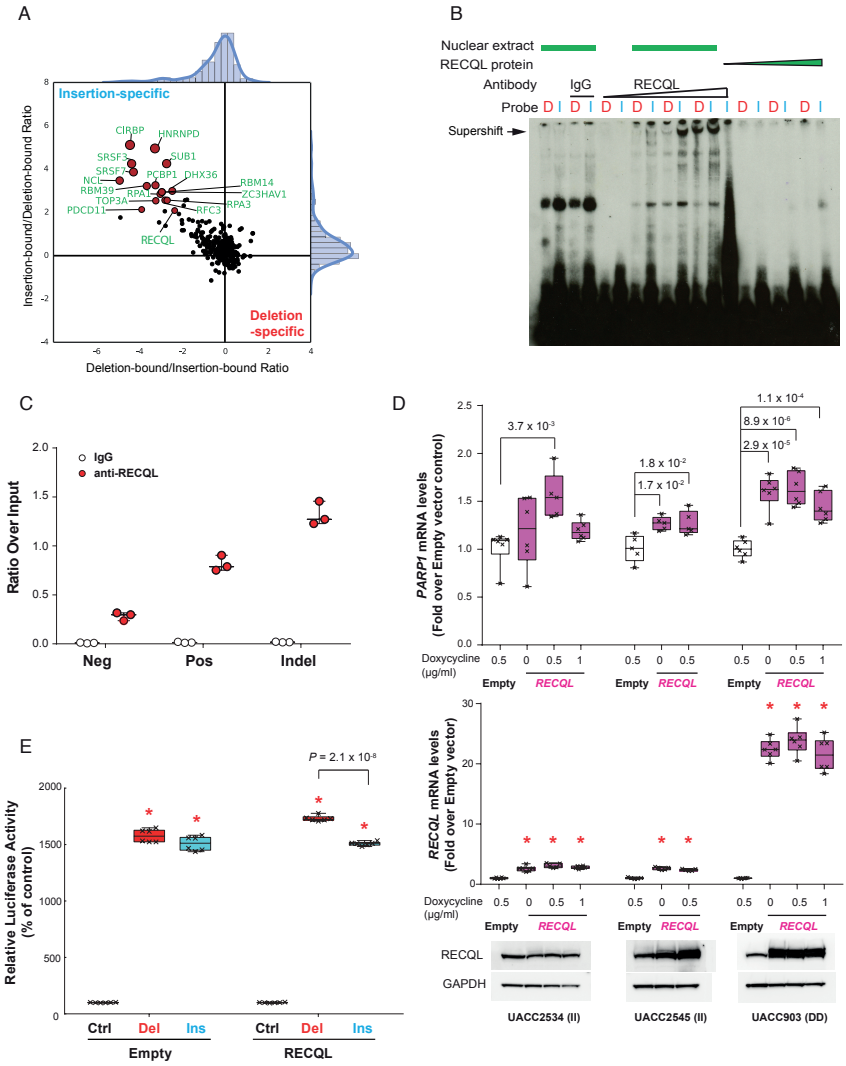


Figure 4. RECQL binds to the insertion allele and mediates allelic expression.

A Insertion allele-specific binding proteins were identified by mass-spectrometry using melanoma cell nuclear extract and biotinylated double-stranded oligos. The ratio of proteins bound to heavy/light-dimethyl labeled probes is plotted on x and y-axis for labeling swapping. Red circles: enriched above the background in both directions. Circle sizes represent relative abundance.

B RECQL EMSA/supershift. D: deletion, I: insertion

C ChIP was performed using anti-RECQL antibody or IgG and melanoma cell chromatin followed by qPCR. DNA quantity was normalized to input DNA for each IP (n=3). Neg: gene desert,

Pos: a known RECQL binding locus, Indel: rs144361550 region. A representative set from four independent experiments is shown.

D RECQL under tetracycline-inducible promoter was expressed in three melanoma cell lines. *PARP1* levels (top) and *RECQL* RNA (middle) and protein (bottom) levels were measured at 48hrs of doxycycline induction (blot images were cropped). Transcript levels are shown as fold over Empty vector after normalizing to *B2M* control (n = 6, 5, and 6 for each cell line).

E Luciferase assays were performed using 905bp deletion (Del) or insertion allele (Ins) constructs with RECQL or Empty vector co-transfection in HEK293FT cells. Renilla-normalized relative luciferase activities were plotted as percent of the minimal promoter control (Ctrl) (n = 6). (C-E) Each graph shows median with range, 75 & 25 percentiles, and each data point. Two-tailed, *t*-test assuming unequal variance for all *P*-values shown. * $P < 0.05$ against Ctrl (E) or Empty (D).

Discussion

In this study, eQTL and ASE analyses suggest *PARP1* as the susceptibility gene underlying the melanoma risk locus on chromosome band 1q42.1. When we evaluated the set of genes in +/- 1Mb of the lead melanoma risk SNP (rs3219090) to account for potential long-range regulation, we observed a highly-reproducible eQTL with *PARP1*, but not with other nearby genes. The correlation between the risk allele and higher levels of *PARP1* expression was highly reproducible across multiple melanoma-relevant tissues, including early-passage melanoma cell lines, melanoma tumors, and human skin biopsies in both eQTL and ASE analyses. While eQTL and ASE analyses cannot completely rule out a potential role for other genes within the larger genomic region surrounding the GWAS peak, these data strongly implicate *PARP1* as functionally mediating melanoma risk at this locus.

While this region is relatively small in size, 65 variants are nonetheless strongly correlated ($r^2>0.6$) with the lead GWAS SNP. To efficiently prioritize functional candidates we took advantage of potential gene regulatory regions annotated in human melanocyte and melanoma samples by the ENCODE and Roadmap Projects. We chose to focus on variants located in most consistently annotated regulatory elements across different individuals and cellular conditions because of the strikingly consistent eQTL and ASE data observed in both melanocytes and melanomas. Subsequent characterization of these candidate variants highlighted a single variant, rs144361550, as a strong functional candidate. Of the variants tested, only rs144361550 demonstrated both allele-specific transcriptional activity and protein binding pattern in a manner consistent with the observed pattern of genotype/expression correlation.

While these data provide support for rs144361550 as a functional melanoma risk variant influencing levels of *PARP1* expression, they nonetheless cannot rule out other variants in this region as also contributing to the observed correlation between *PARP1* levels and genotype.

Our unbiased approach using quantitative mass-spectrometry identified RECQL as a protein binding allele-preferentially to rs144361550. Importantly, RECQL binding to rs144361550 does not appear to be driven by sequence specificity but rather by DNA secondary structure. While genomic sequence encompassing rs144361550 suggested G4-forming potential (Supplementary Table 11), which might explain a regulatory role[36], our *in vitro* assays failed to provide definitive evidence for G4 structure either by insertion or deletion allele. However, formation of another differential structural motif, such as a transient hairpin structure (formed by single-stranded sequences, Supplementary Table 13) or a locally perturbed double-helix structure at the hexanucleotide repeat domain[37], inducing DNA bending and serving as a recognition motif for allele-specific protein binding[38], cannot be excluded. As such, a more detailed molecular and structural analysis of the interaction between the rs144361550 insertion and deletion alleles, in both a dsDNA and ssDNA context, will be an important topic for future research.

References

1 Amos, C. I. *et al.* Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Hum Mol Genet* **20**, 5012-5023, doi:10.1093/hmg/ddr415 (2011).

2 Barrett, J. H. *et al.* Genome-wide association study identifies three new melanoma susceptibility loci. *Nature genetics* **43**, 1108-1113, doi:10.1038/ng.959 (2011).

3 Barrett, J. H. *et al.* Fine mapping of genetic susceptibility loci for melanoma reveals a mixture of single variant and multiple variant regions. *Int J Cancer* **136**, 1351-1360, doi:10.1002/ijc.29099 (2015).

4 Bishop, D. T. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet* **41**, 920-925, doi:10.1038/ng.411 (2009).

5 Brown, K. M. *et al.* Common sequence variants on 20q11.22 confer melanoma susceptibility. *Nature genetics* **40**, 838-840, doi:10.1038/ng.163 (2008).

6 Iles, M. M. *et al.* A variant in FTO shows association with melanoma risk not due to BMI. *Nat Genet* **45**, 428-432, 432e421, doi:10.1038/ng.2571 (2013).

7 Law, M. H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat Genet* **47**, 987-995, doi:10.1038/ng.3373 (2015).

8 Macgregor, S. *et al.* Genome-wide association study identifies a new melanoma susceptibility locus at 1q21.3. *Nature genetics* **43**, 1114-1118, doi:10.1038/ng.958 (2011).

9 Rafnar, T. *et al.* Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet* **41**, 221-227, doi:10.1038/ng.296 (2009).

10 Pena-Chilet, M. *et al.* Genetic variants in PARP1 (rs3219090) and IRF4 (rs12203592) genes associated with melanoma susceptibility in a Spanish population. *BMC Cancer* **13**, 160, doi:10.1186/1471-2407-13-160 (2013).

11 Davies, J. R. *et al.* Inherited variation in the PARP1 gene and survival from melanoma. *Int J Cancer* **135**, 1625-1633, doi:10.1002/ijc.28796 (2014).

12 Law, M. H. *et al.* PARP1 polymorphisms play opposing roles in melanoma occurrence and survival. *Int J Cancer* **136**, 2488-2489, doi:10.1002/ijc.29280 (2015).

13 Krishnakumar, R. & Kraus, W. L. The PARP side of the nucleus: molecular actions, physiological outcomes, and clinical targets. *Mol Cell* **39**, 8-24, doi:10.1016/j.molcel.2010.06.017 (2010).

14 Woodhouse, B. C. & Dianov, G. L. Poly ADP-ribose polymerase-1: an international molecule of mystery. *DNA Repair (Amst)* **7**, 1077-1086, doi:10.1016/j.dnarep.2008.03.009 (2008).

15 Huber, A., Bai, P., de Murcia, J. M. & de Murcia, G. PARP-1, PARP-2 and ATM in the DNA damage response: functional synergy in mouse development. *DNA Repair (Amst)* **3**, 1103-1108, doi:10.1016/j.dnarep.2004.06.002 (2004).

16 Bouchard, V. J., Rouleau, M. & Poirier, G. G. PARP-1, a determinant of cell survival in response to DNA damage. *Exp Hematol* **31**, 446-454 (2003).

17 Maynard, S., Schurman, S. H., Harboe, C., de Souza-Pinto, N. C. & Bohr, V. A. Base excision repair of oxidative DNA damage and association with cancer and aging. *Carcinogenesis* **30**, 2-10, doi:10.1093/carcin/bgn250 (2009).

18 Swindall, A. F., Stanley, J. A. & Yang, E. S. PARP-1: Friend or Foe of DNA Damage and Repair in Tumorigenesis? *Cancers (Basel)* **5**, 943-958, doi:10.3390/cancers5030943 (2013).

19 Urabe, K. *et al.* The inherent cytotoxicity of melanin precursors: a revision. *Biochim Biophys Acta* **1221**, 272-278 (1994).

20 Kuilman, T., Michaloglou, C., Mooi, W. J. & Peeper, D. S. The essence of senescence. *Genes Dev* **24**, 2463-2479, doi:10.1101/gad.1971610 (2010).

21 Leikam, C., Hufnagel, A., Schartl, M. & Meierjohann, S. Oncogene activation in melanocytes links reactive oxygen to multinucleated phenotype and senescence. *Oncogene* **27**, 7070-7082, doi:10.1038/onc.2008.323 (2008).

22 Kraus, W. L. Transcriptional control by PARP-1: chromatin modulation, enhancer-binding, coregulation, and insulation. *Curr Opin Cell Biol* **20**, 294-302, doi:10.1016/j.ceb.2008.03.006 (2008).

23 Schiewer, M. J. & Knudsen, K. E. Transcriptional roles of PARP1 in cancer. *Mol Cancer Res* **12**, 1069-1080, doi:10.1158/1541-7786.MCR-13-0672 (2014).

24 Hassa, P. O. & Hottiger, M. O. The functional role of poly(ADP-ribose)polymerase 1 as novel coactivator of NF-kappaB in inflammatory disorders. *Cell Mol Life Sci* **59**, 1534-1553 (2002).

25 Cervellera, M. N. & Sala, A. Poly(ADP-ribose) polymerase is a B-MYB coactivator. *J Biol Chem* **275**, 10692-10696 (2000).

26 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

27 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

28 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

29 Butter, F. *et al.* Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet* **8**, e1002982, doi:10.1371/journal.pgen.1002982 (2012).

30 Hubner, N. C., Nguyen, L. N., Hornig, N. C. & Stunnenberg, H. G. A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. *J Proteome Res* **14**, 1315-1329, doi:10.1021/pr5009515 (2015).

31 Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res* **11**, 1475-1489 (1983).

32 Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* **4**, 484-494, doi:10.1038/nprot.2009.21 (2009).

33 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

34 McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem* **84**, 7469-7478, doi:10.1021/ac301572t (2012).

35 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

36 Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-780, doi:10.1038/nrg3296 (2012).

37 Stefl, R. *et al.* A-like guanine-guanine stacking in the aqueous DNA duplex of d(GGGGCCCC). *J Mol Biol* **307**, 513-524, doi:10.1006/jmbi.2001.4484 (2001).

38 Lu, X. J., Shakked, Z. & Olson, W. K. A-form conformational motifs in ligand-bound DNA structures. *J Mol Biol* **300**, 819-840, doi:10.1006/jmbi.2000.3690 (2000).

SUPPLEMENTARY NOTE

Supplementary Results

Evaluation of alternative splicing as a potential mechanism of *PARP1* regulation

We considered other potential mechanisms by which the two alleles of rs144361550 may influence *PARP1* levels. Given the location of rs144361550 in an intron, approximately 100bp downstream of the first exon-intron junction, we investigated the possibility of allelic *PARP1* regulation via alternative splicing. Sequence-based prediction suggested that neither the insertion nor deletion alleles create cryptic splice donor, branch point, or acceptor sites (http://splice.uwo.ca/)[1,2]. In addition, RNAseq data from 15 early-passage melanoma cell lines used in this study, as well as three independent primary melanocyte cultures, did not detect novel or cryptic splice forms of *PARP1* transcript including an unspliced first intron sequence, and a larger qPCR-based analysis of 57 early-passage melanoma cell lines and six primary melanocyte cultures did not show a statistically significant association between the melanoma risk SNP rs3219090 and any specific alternative *PARP1* transcript (data not shown). Together, these data do not support altered splicing as a likely functional mechanism.

Prioritization of candidate functional variants

We prioritized 65 variants that are highly correlated with the lead SNP as candidate functional variants (r^2>0.6 with lead SNPs from the discovery or meta-analysis lead SNPs[3,4], rs3219090 and rs1858550, respectively; LD based on 1KG phase3, EUR and CEU). This set of 65 variants included all that were found to be associated with melanoma within four orders of magnitude *P*-value of the lead SNP from the latest meta-analysis[4] (n = 51 SNPs, **Supplementary Table 5-6**). Given the absence of amino acid-changing *PARP1* variants within this set of candidates as well as the considerable evidence for allelic differences in *PARP1* expression levels, we focused on those located within annotated melanocyte- or melanoma specific *cis*-regulatory elements[5,6]. Considering that accessible chromatin regions annotated using DNase I hypersensitivity sequencing (DHS) data are one of the most inclusive predictors of different classes of tissue-specific *cis*-regulatory elements[7], we turned to DHS data

generated from three independent cultures of primary human melanocytes, available through the ENCODE[5] and Roadmap projects[6], as a primary predictor (**Supplementary Fig. 2**). As a secondary predictor, we examined DHS from two melanoma cell lines, as well as recently published open chromatin data using Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) from 11 melanoma cultures and cell lines[8] (**Supplementary Fig. 3**). We considered the strongest candidates to be those that are situated in regions of open chromatin in both primary melanocyte and melanoma cultures (open chromatin in all three melanocytes and >50% of melanoma cultures), and identified four such variants (**Supplementary Figs. 2-3**). All four of the most strongly supported variants are situated at the center of melanocyte DHS peaks as well as within regions harboring promoter or enhancer histone marks (H3K4me1, H3K4me3, or H3K27ac) in the majority of melanocyte/melanoma cultures (**Supplementary Table 6**). Based on these data, we proceeded with functional characterization of these four candidates (**Supplementary Table 6, Supplementary Fig. 2**).

Direct genotyping of rs144361550 and LD assessment

Notably, one of the top four candidates, rs144361550, is a six base-pair insertion/deletion (indel) within a string of GGGCCC repeat units at the beginning of the first intron of *PARP1*. This variant lies over prominent melanocyte epigenetic marks, including H3K27Ac, H3K4me3, and DNAseI hypersensitive peaks, as well as ENCODE ChIP-seq peaks for multiple transcription factors (**Fig.2, Supplementary Fig. 4**). Consistent with its annotation as an active transcriptional start site (TSS) by the Epigenome RoadMap Project, this region displays a chromatin modification signature consistent with a promoter region, where H3K4Me3 and H3K27Ac peaks are overlapping while H3K4me1 signal is diminished. LD estimates between rs144361550 and the lead SNP (rs3219090) within this locus varied between 1KG phase 1 ($r^2 = 0.67$), which served as the imputation reference for meta-analysis datasets[4], and 1KG phase 3 ($r^2 = 0.95$). Given this disparity, we directly genotyped this indel for 85 HapMap CEU individuals using a fluorescence-based fragment length polymorphism assay on a capillary sequencer (**Supplementary Fig. 7, Supplementary Table 8**). Of 35 individuals included in both HapMap and 1KG phase 1, three individuals showed discordant genotypes for rs144361550 (~9% estimated error rate); data from HapMap samples were consistent with Mendelian errors introduced only

from 1KG phase1 genotypes (**Supplementary Fig. 8**). We then further assessed LD between rs144361550 and rs3219090 by directly genotyping rs144361550 in a set of 745 healthy individuals of European descent[9]; correlation between these two variants was consistent with that observed in 1KG phase 3 (r^2=0.94; **Supplementary Table 9**), where the deletion allele is phased with the rs3219090 risk allele G.

In silico and *in vitro* tests of G-quadruplex DNA structure formation on rs144361550

In an attempt to understand the seemingly unconventional mechanism of allele-specific binding of RECQL to rs144361550, we performed a set of *in vitro* assays assessing the nature of the DNA secondary structure driving the indel allelic difference. Based on the evidence that (i) RECQL is a helicase as opposed to transcription factor recognizing and binding DNA sequence motifs, (ii) 41% of insertion-binding proteins from the mass-spec including RECQL are also known to bind alternative nucleotide secondary structures including G-quadruplex (G4) (**Supplementary Table 10**)[10-15], and (iii) sequence prediction analyses suggesting allelic differences in G4-forming potential for sequences encompassing rs144361550 (**Supplementary Table 11**), we tested if G4 structure can be formed by insertion or deletion alleles, *in vitro*. A series of biophysical studies including circular dichroism (CD) spectrometry (**Supplementary Fig. 11 and 12**) and thermal difference UV spectroscopy (TDS, **Supplementary Fig. 13**) suggested that a G4 structure could be induced in the presence of a strong G4 ligand, PhenDC3, for both insertion and deletion alleles but without allelic discrimination. These results failed to demonstrate a definitive G4 structure formed either by insertion or deletion allele *in vitro* and are also consistent with the lack of validation for the other six G4-binding proteins from mass-spec by antibody-supershifts, suggesting a RECQL-specific allelic binding mechanism.

Supplementary Methods

qPCR analysis of *PARP1* in early passage melanoma cell lines

The same RNA used for expression array analysis was used for cDNA generation for qPCR analysis. Expression levels of four control genes were measured in all 57 cell lines, and two most stable genes selected from Biogazelle qbasePLUS analysis (*ACTB*, *PPIA*) were used for duplex Taqman assays with VIC and FAM labeled probe sets (Taqman probe ID Hs00242302). Geometric means of control gene Ct values from four qPCR replicates (2 x *ACTB* and 2 x *PPIA*) were subtracted from geometric means of *PARP1* Ct values to generate mean ΔCt values. Mean ΔCt from replicates were treated as single data points. After assessing normal distribution by Shapiro-Wilk test, analysis of covariance was performed for mean ΔCt using genotype as a category. Only the samples whose genotype (rs3219090) was directly typed by array (n=53) were included in the analysis.

eQTL and ASE analysis from TCGA and GTEx RNAseq datasets

For TCGA melanomas eQTL analysis, transcript levels, rs3219090 genotypes, and Affymetrix SNP6 regional copy numbers were obtained from TCGA data version 2015_11_01 (dbGAP Accession: phs000178.v9.p8) through Firebrowse beta version (http://firebrowse.org/) and the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). RNAseq expression data was extracted as "RNA Seq V2 RSEM" from level 3 file "rem.genes.normalized_ results" and rs3219090 genotype data was generated from level 2 "birdseed. data.txt" with filter set at confidence threshold <0.05. SNP6 copy numbers were averaged across the genomic region of each gene to obtain gene-based copy number. A total of 409 melanoma samples from unique individuals had direct genotype, expression, and copy number information available and were used for the analysis. A total of 16 genes, including *PARP1*, were located within the 2 Mb region surrounding rs3219090 had genotype, expression, and copy number data available and were used for the analysis. The Matrix eQTL package (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/) was used for eQTL analysis, using a linear model considering an additive model for genotypes and using gene-specific copy number as a covariate. To incorporate gene-specific copy numbers as a covariate, Matrix eQTL analysis

was run separately for each gene. GTEx eQTL analysis was performed on GTEx portal (http://www.gtexportal.org/home/testyourown; 12/02/2015,data source: GTEx Analysis Release V6, dbGaP Accession phs000424.v6.p1) using the Test Your Own eQTL function for 20 genes within +/- 1Mb of rs3219090 (RefSeq Genes, hg19) that had detectable expression data available. For ASE, we included 48 TCGA melanoma samples from unique individuals that are both heterozygous for rs1805414 and rs3219090, as well as copy-neutral for the *PARP1* genomic region by GISTIC 2.0 (www.broadinstitute.org/cancer/cga/gistic, obtained through cBio portal; http://www.cbioportal.org/). rs3219090 genotypes were obtained from Affymetrix SNP6 genotype data, while allele-specific reads for assessing rs1805414 genotype were obtained from exome sequencing data. Allele-specific *PARP1* transcript levels were obtained from RNAseq reads (TCGA, dbGAP Accession: phs000178.v9.p8). For accurate genotype assignment, only the samples with >=5 reads for each allele were accepted as heterozygotes for rs1805414. The mapped numbers of RNAseq reads encompassing the variant for each allele were used to calculate an allelic ratio. Only those samples with >= 10 reads for each allele were used for the analysis. A two-tailed Wilcoxon signed rank test was used to assess allelic imbalance. GTEx ASE analysis was performed in a similar manner with the following difference. Samples included in the analysis were heterozygous for rs1805414 based on genotype imputation and also heterozygous for rs3219090 by genotyping (n = 139, sun-exposed skin; n = 69, not sun-exposed; GTEx Analysis Release V6, dbGaP Accession phs000424.v6.p1,12/17/2015). All samples have >= 10 total reads and >= 3 reads for each allele.

Re-genotyping of rs144361550

Two different amplicons (156bp and 240bp) encompassing rs144361550 were amplified from genomic DNA of 85 HapMap CEU individuals and 745 healthy individuals of European descent from the NCI-DCEG imputation reference panel[9] using respective primer sets labeled with different fluorescent dyes (6-FAM for 156 bp and VIC for 240 bp, primer sequences listed in **Supplementary Table 14**). PCR products were purified and subsequently injected into an ABI3730XL with the Gene Scan 500LIZ size standard. Genotypes were automatically called using GeneMapper (Applied Biosystems/Thermo Fisher Scientific) and manually confirmed using PeakScanner v2. A

subset of samples with ambiguous calls were validated by Sanger sequencing followed by analysis using Mutation Analyzer (SoftGenetics, State College, PA) using strand separation function for mixed sequences of heterozygous samples.

Circular dichroism (CD) and thermal difference spectra (TDS)

Lyophilized, HPLC-purified oligonucleotides (sequences: cf. **Supplementary Table 14**) were purchased from MWG Eurofins and dissolved in milliQ water to a strand concentration of 200 μM. Single-stranded and double-stranded samples for CD and TDS were prepared to give final total strand concentration of 5 μM in 10 mM Li cacodylate buffer (pH 7.2) supplemented with 100 mM LiCl (Li^+ conditions), 100 mM KCl (K^+ conditions) or 100 mM KCl + 10 μM PhenDC3, a G4-stabilizing ligand[16]. Samples were heated to 95 °C for 5 min, slowly cooled to ambient temperature, and kept at +4 °C overnight, in order to induce formation of thermodynamically stable secondary structures. CD spectra were recorded with a JASCO J-710 spectropolarimeter equipped with a Peltier temperature controller, using quartz cells with a path length of 10 mm; the scans were recorded at 20 °C from 210 to 330 nm using the following parameters: data pitch, 0.5 nm; bandwidth, 2 nm; response, 2 s; scan speed, 50 nm min^{i1}; the scans are the result of four accumulations. The CD spectra were blank-subtracted and converted to molar dichroic absorption (Δε, cm^{-1} M^{-1}) based on total nucleoside concentration (Equation S1):

$$\Delta\varepsilon_{CD} = \Theta / 32980 \times c \times n \times \ell) \quad \text{(S1)}$$

where Θ is the ellipticity (millidegrees), c is the strand concentration in sample (M), n is oligonucleotide length (bases), and ℓ is the path length (cm).

Absorption spectra were obtained with an Agilent Cary 300 Bio spectrophotometer equipped with a temperature controller in quartz cells (path length: 1 cm), using samples containing 5 μM DNA (strand concentration) in K^+-containing buffer as mentioned above. Absorption spectra were recorded at 20 °C and then at 90 °C, and molar extinction coefficient differences (TDS) were calculated using Equation S2:

$$\Delta\varepsilon TDS = [A(90\ °C) - A(20\ °C)] / (c \times n \times \ell) \quad \text{(S2)}$$

where A is absorbance at a given temperature, and c, n, and ℓ are as above.

Bisulfite sequencing

Genomic DNA were purified from melanocytes using the DNeasy Blood and Tissue Kit from Qiagen, and bisulfite conversion was performed using the EZ DNA methylation-Direct kit from ZYMO research Corp, following the manufacturer's instructions. PCR-amplified bisulfite-converted DNAs were then sequenced on a 3730xl DNA Analyzer (ABI). The sequence of PCR primers are listed in **Supplementary Table 14**. Methylated CpG control DNA was provided by the kit.

Luciferase assays for *MITF* promoter

Three luciferase constructs were generated containing different lengths of the *MITF-M* promoter. Approximately 2.2 Kb 5' region of *MITF-M* promoter was sub-cloned into pGL4.23 luciferase vector (Promega) from p*MITF*-2256 (a gift from Dr.

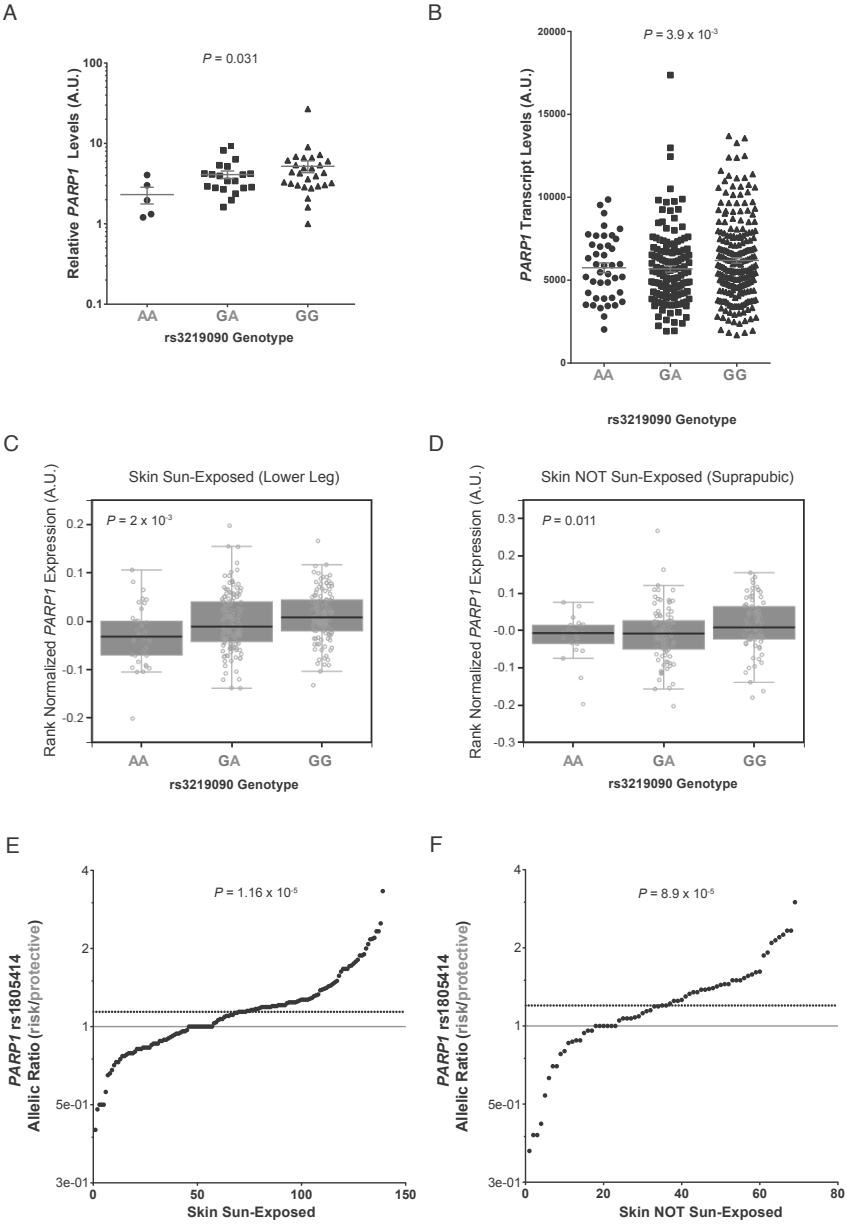
William Pavan, NHGRI)[17] to make MITF.2200. A short 382bp region of the *MITF-M* promoter was sub-cloned into pGL4.23 from p*MITF*-382 (a gift from Dr. David Fisher, Dana-Farber Cancer Institute)[17] to make MITF.382. A 674bp fragment 5' of *MITF-M* promoter was PCR amplified from p*MITF*-2256 and cloned into pGL4.23 to make MITF.674. Primer sequences for sub-cloning are listed in **Supplementary Table 14**. pGL4.23 constructs were then co-transfected with pGL4.74 (Renilla luciferase) into melanocytes infected with either *PARP1*shRNA or *PARP1* expression vector by electroporation with Lonza Amaxa P2 kit and protocol CA-137 (Lonza). Cells were collected 24hr following transfection and luciferase activity was measured using Dual-Luciferase reporter system (Promega) on GLOMAX multi detection system (Promega).

Prediction of PARP1 binding sites on *MITF-M* promoter

FASTA file of the *MITF* promoter sequence was downloaded from UCSC genome browser (hg19), and loaded to CLC Genomics Workbench (version 7.5). We used the TRANSFAC TFBS (BIOBASE professional database) Plugin (version 1.1) on CLC Genome Workbench and filtered the result by setting "Set Matrix similarity cut-off" as "only high-quality matrices" ("set matrix similarity cut-off 0.95 and set core similarity cut-off 0.99"). Predicted PARP1 binding sites were extracted from the result list.

Expression correlation analysis

PARP1 and *MITF* transcript levels, and rs3219090 genotypes of 409 cutaneous melanomas were obtained from TCGA (same set of samples that were used for eQTL). GISTIC copy numbers of *MITF* were accessed through cBioPortal[18] and 189 neutral copy samples were assessed separately for correlation. Pearson correlation analysis was performed to obtain correlation coefficient (r). For 59 early passage melanoma cell lines, expression levels of *PARP1*, *MITF*, and five *MITF* target genes (*CDK2, TBX2, RAB27A, EDNRB*, and *MC1R*) were obtained from microarray data in the same way as *PARP1* eQTL analysis. Normalized *MITF* levels of 59 cell lines were plotted, and cell lines were divided into high- and low-*MITF* subgroups based on a distinct delineation point. Genomic copy number of *MITF* region was also assessed using genotyping data, with no cell line displayed high-level amplification of the *MITF* gene. Pearson correlation analysis was performed to obtain correlation coefficient (r), or linear regression was used when *MITF* copy number was used as a covariate.



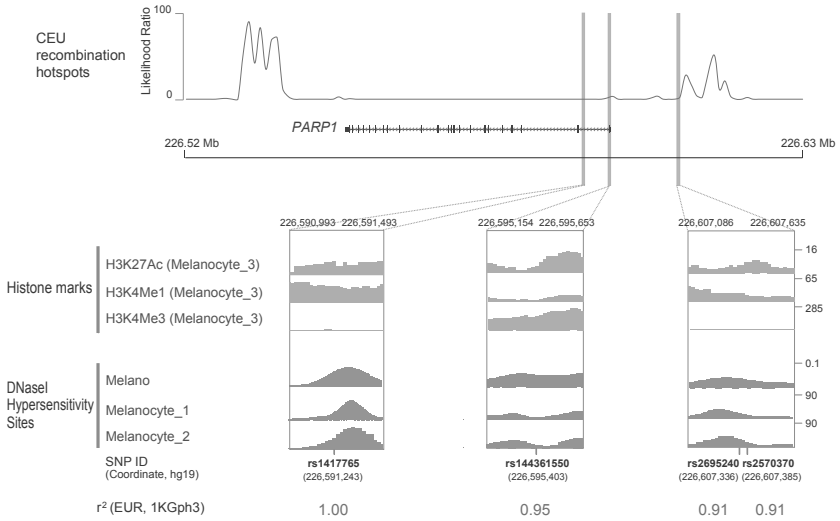
Supplementary Figure 1. *PARP1* eQTL and allele-specific expression analysis of melanoma cell lines, TCGA melanomas, and GTEx skin tissues.

A *PARP1* eQTL validation by qPCR in low passage human melanoma cell lines. *PARP1* transcript levels in 53 melanoma cell lines were measured by Taqman qPCR. *PARP1* levels were normalized against those of two control genes (*ACTB* and *PPIA*) and relative fold difference to the lowest sample is plotted using the ΔΔCt method ($P = 0.031$, linear regression). G is the risk allele and A the protective allele of rs3219090. A.U. ; arbitrary unit, horizontal lines represent mean value and error bars s.e.m.

B *PARP1* expression levels from RNA sequencing of 409 melanomas as a part of the TCGA skin melanoma project were plotted for each genotype of rs3219090 ($P = 3.9 \times 10^{-3}$, linear regression using gene-specific regional copy number as a covariate). Horizontal lines represent mean value and error bars s.e.m.

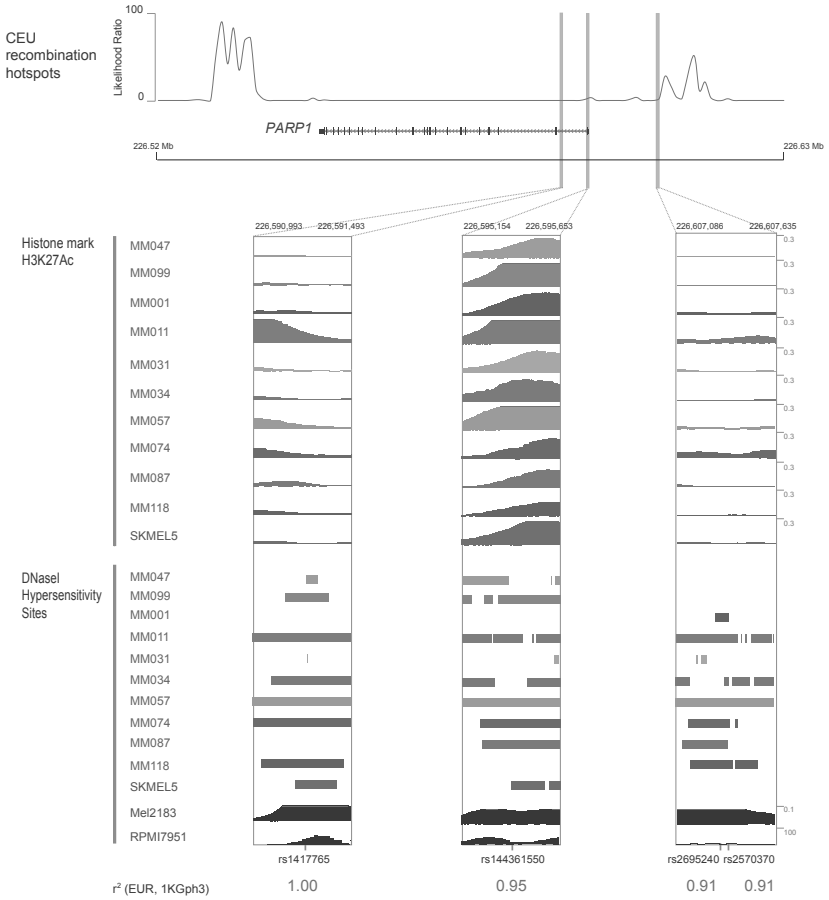
C,D *PARP1* eQTL analysis in GTEx skin tissues. eQTL analysis of *PARP1* levels and genotype of rs3219090 was performed on 302 sun-exposed skin samples ($P = 2 \times 10^{-3}$, linear regression), as well as 196 skin samples from non-sun-exposed skin ($P = 0.01$, linear regression) using the GTEx portal (GTEx Analysis Release V6; dbGaP Accession: phs000424.v6.p1; http://www.gtexportal.org/home/testyourown).

E, F Allelic ratios of *PARP1* transcripts were measured using data from (E) 139 sun-exposed or (F) 69 non-sun-exposed skin samples that are heterozygous for both rs3219090 and rs1805414 (coding surrogate SNP of rs3219090) from the GTEx database. The mapped numbers of RNAseq reads encompassing each allele of rs1805414 variant were used for calculating allelic ratios ($P = 1.16 \times 10^{-5}$, sun-exposed; $P = 8.9 \times 10^{-5}$, non-sun-exposed; two-tailed Wilcoxon signed rank test). Solid line marks 1:1 ratio, and dashed line represents the median ratio.



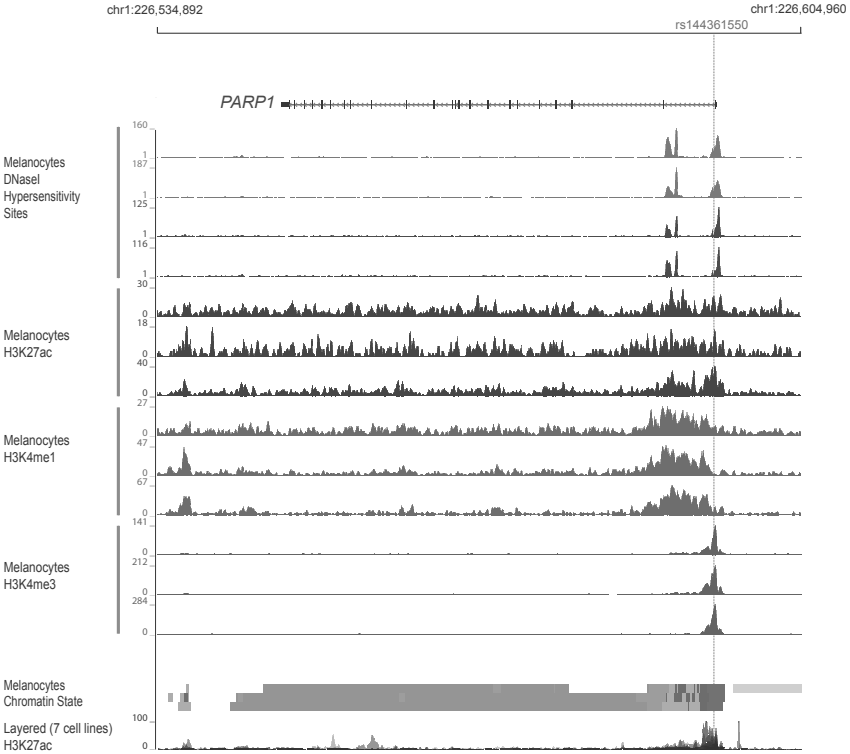
Supplementary Figure 2. Functional annotation and prioritization of potentially gene-regulatory candidate sequence variants based on melanocyte- and melanoma-specific epigenetic data.

Prioritization of candidate functional variants was based on LD with the lead GWAS SNPs (rs3219090 and rs1858550; $r^2 > 0.6$) as well as location within potentially gene regulatory regions in human melanocyte and melanoma samples. All tracks for melanocyte DNaseI Hypersensitivity Site (DHS) and histone mark data were obtained from ENCODE and Roadmap Projects through UCSC genome browser. Histone mark ChIP-seq signals (H3K4Me1, H3K4Me3, and H3K27Ac) are shown for a representative individual (melanocyte_3) among three individuals analyzed by the Roadmap Project. DNaseI hypersensitivity signals are shown for one individual from ENCODE project (Melano) and two individuals from Roadmap project (Melanocyte_1, Melanocyte_2). The scale of each track is uniformly set throughout the region of the *PARP1* gene to cover the highest peaks, with 0 as the baseline (see online methods for details of each track). Regions spanning 250bp upstream and downstream of each candidate SNP are enlarged in boxes. The genomic scale on chromosome 1 is following the hg19 human genome build and presented in bp unless it is shown as Mb. r^2 of each SNP with rs3219090 was obtained from 1000 Genomes Phase 3 EUR (EUR, 1KGph3) population. Recombination hotspots for 1000 genomes phase1 v3 CEU population are presented as log likelihood ratio.



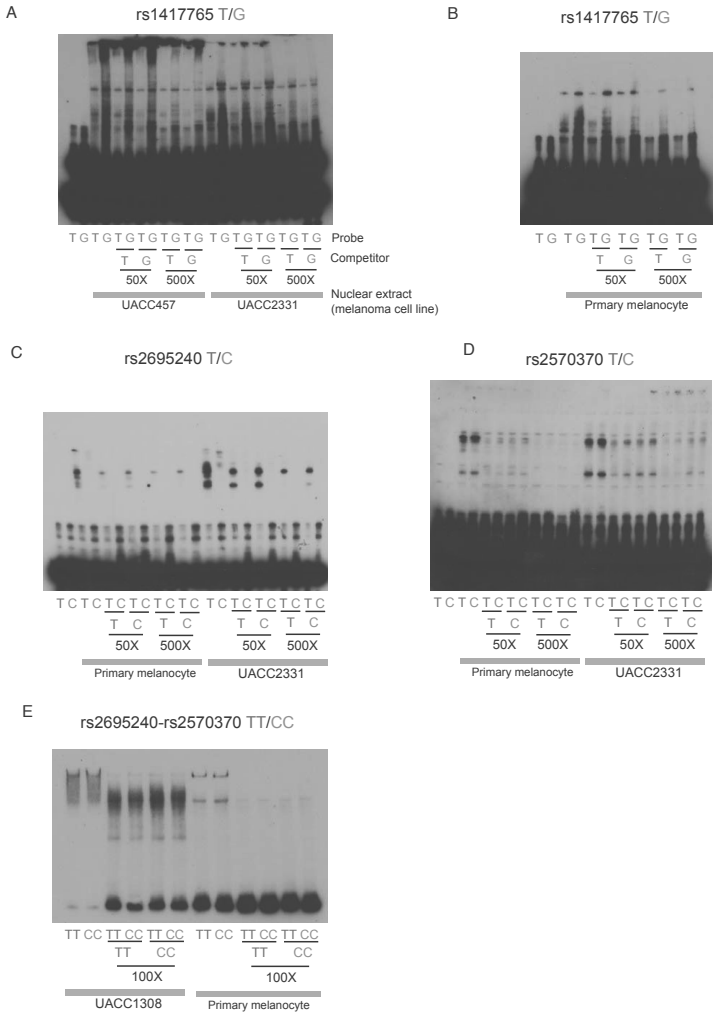
Supplementary Figure 3. Functional annotation of potentially gene- regulatory sequence variants at the *PARP1* locus.

Annotations of four strong candidates are shown for both melanoma open chromatin and histone marks. All tracks were obtained from Melanoma Epigenome Project [8] and the ENCODE Project through the UCSC genome browser. Histone marks (H3K27Ac) are shown for 11 samples from Melanoma Epigenome. Scales of H3K27Ac ChIP-seq signal are uniformly set throughout the genome (0 to 0.3). Peaks are shown for 11 samples from Melanoma Epignome and two melanoma cell lines from ENCODE project (Mel2183 and RPMI7951). Open chromatin peaks from FAIRE-seq are presented as interval of peaks except for Mel2183 and RPMI7951 for which uniform scale of DNaseI hypersensitivity signal was used throughout the *PARP1* gene (0 to 0.1 and 0 to 100, respectively). See online method for details of each track. Regions spanning 250bp upstream and downstream of each candidate SNP are enlarged. Genomic scale on chromosome 1 follows the hg19 human genome build and presented in bp unless otherwise shown as Mb. r^2 of each SNP with rs3219090 was obtained from 1000 Genomes Phase3 (1KGph3) EUR population. Recombination hotspots for 1000 Genomes Phase1 v3 CEU population are presented as log likelihood ratio.



Supplementary Figure 4. Epigenetic annotation of the region encompassing *PARP1* in primary melanocytes.

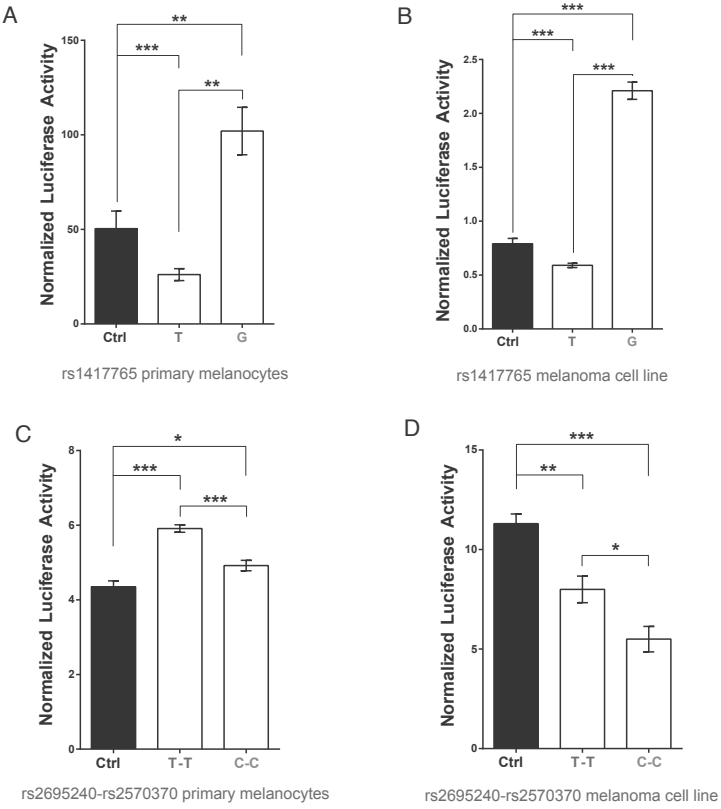
A zoomed out view of histone modification and DNaseI hypersensitivity sites (DHS) in primary melanocytes are shown for *PARP1* region. Red dashed vertical line indicates the position of rs144361550 overlapping histone marks and DHS. Genomic positions are based on hg19. Melanocyte DNaseI Hypersensitivity Sites are shown for two experimental replicates of two individuals. H3K4me1 and H3K4me3 signals are from three individuals, and H3K27ac traces are shown for two individuals including an experimental replicate. Melanocyte Chromatin States are from Chromatin Primary Core Marks Segmentation by HMM from Roadmap Project. Red: TSS; yellow: Enhancers; Green: Transcription. All tracks were obtained from ENCODE and Roadmap projects through UCSC Genome browser.



Supplementary Figure 5. EMSAs of three top functional candidate SNPs rs1417765, rs2695240, and rs2570370.

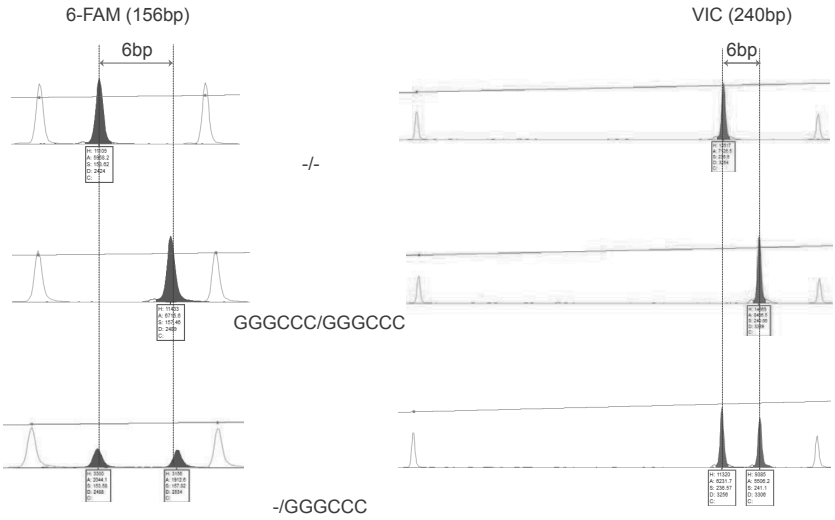
A-D EMSAs were performed using biotin-labeled double-stranded DNA probes encompassing 10 bp to either side of each SNP with two different alleles. Nuclear extracts from melanoma cell lines UACC457, UACC1308, or UACC2331, and primary human melanocytes were used for each SNP. 50X, 100X or 500X molar excess of unlabeled competitors were added in specified lanes for competition. Risk alleles are shown first (T allele), in red, and protective alleles second (G or C allele) in green.

E Since rs2695240 and rs2470370 are 50bp apart and in perfect LD, two SNPs were also tested as haplotypes using 186bp probes encompassing both SNPs. Probes were PCR-amplified using biotin labeled primers (primer sequences in **Supplementary Table 14**).



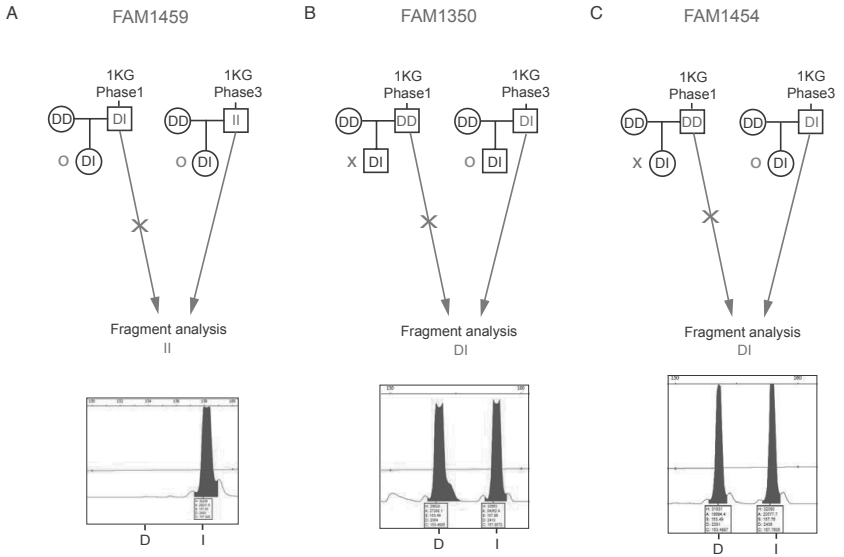
Supplementary Figure 6. Luciferase assays for rs1417765 and rs2695240.

Sequences encompassing rs1417765 (557bp) or rs2695240-rs2570370 (711bp) were cloned 5' of minimal promoter of pGL4.23 vector and transfected into (A,C) primary melanocytes or (B) melanoma cell lines UACC1308 and (D) UACC2331. Luciferase activity was measured 24hrs after transfection and normalized against Renilla luciferase activity (shown as mean with s.e.m.). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ (n=6, two-tailed, *t*-test assuming unequal variance). Risk alleles are shown first (T allele), in red, and protective alleles second (G or C allele) in green. Since rs2695240 and rs2570370 are 50bp apart and in perfect LD, two SNPs were tested as a haplotype.



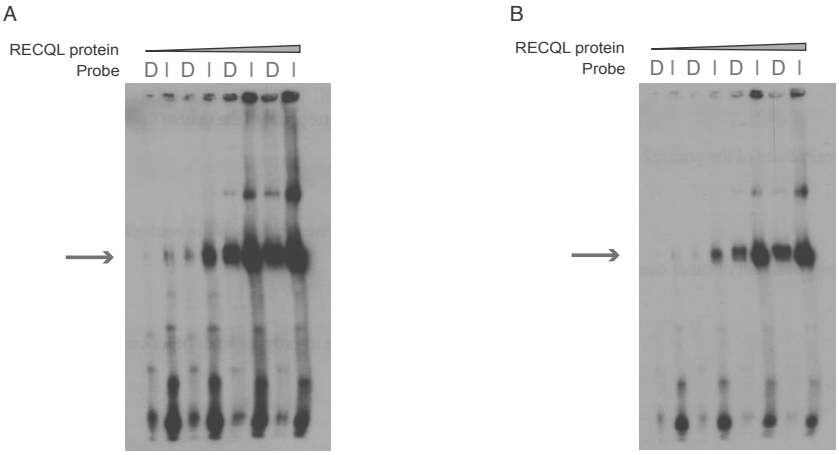
Supplementary Figure 7. Representative capillary electrophoresis profiles of fragment analysis for rs144361550.

Each panel shows an example trace for each of three genotypes (-/-: deletion/deletion, GGGCCC/GGGCCC: insertion/insertion, and -/GGGCCC: deletion/insertion). Orange traces represent the sizing ladder, while blue and green peaks represent 6-FAM and VIC signal, respectively.



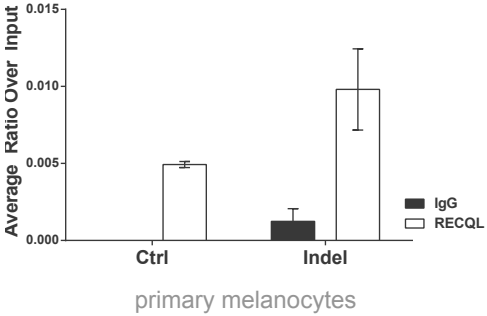
Supplementary Figure 8. Hapmap CEU individuals with discordant genotypes between 1KG Phases 1 and 3 were resolved by fragment analysis.

Three individuals from three Hapmap CEU trios (families 1459, 1350, and 1454) with discordant genotypes of rs144361550 between 1000 Genomes Phase1 (1KG Phase1) and Phase 3 (1KG Phase3) are shown twice, with genotypes from each 1KG Phase. The remaining two members of the trios were genotyped by fragment analysis and resulting genotypes are shown in black. Fragment analysis genotypes and traces for discordant individuals are shown at the bottom and in the boxes: D, deletion allele; I, insertion allele. A red 'X' next to each progeny of trios denotes Mendelian error while green 'O' denotes no Mendelian error. All three discordant genotypes were resolved by fragment analysis and validated 1KG Phase3 genotypes. Trio structures and genotype comparisons are shown in **Supplementary Table 8**.



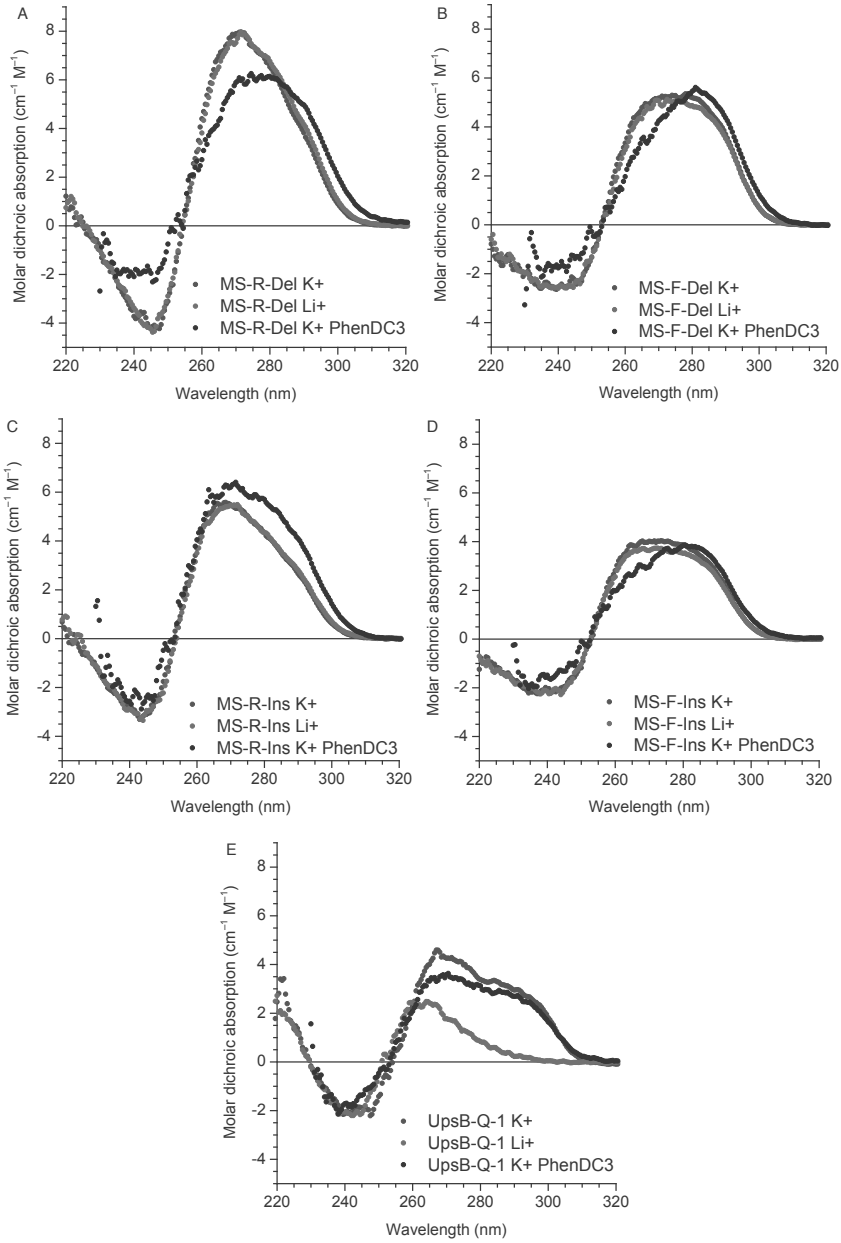
Supplementary Figure 9. RECQL preferentially binds to the rs144361550 insertion allele.

EMSA was performed using 100-800ng purified recombinant RECQL protein. Arrow indicates RECQL-probe complex. D: double-stranded 22bp deletion probe (risk), I: double-stranded 28bp insertion probe (protective). (A) long exposure, (B) short exposure.



Supplementary Figure 10. RECQL binds the genomic region encompassing rs144361550 in primary melanocytes.

Chromatin immunoprecipitation was performed using anti-RECQL antibody or normal IgG and sheared chromatin from 1% formaldehyde-fixed primary melanocytes. DNA was isolated from pulled-down chromatin and analyzed by qPCR. DNA quantity was normalized by taking ratio over input DNA for each IP. qPCR primers were designed to recognize either a known RECQL binding locus (Ctrl) or the region encompassing rs144361550 (indel). The average ratio from two independent experiments is presented (mean with s.e.m.).

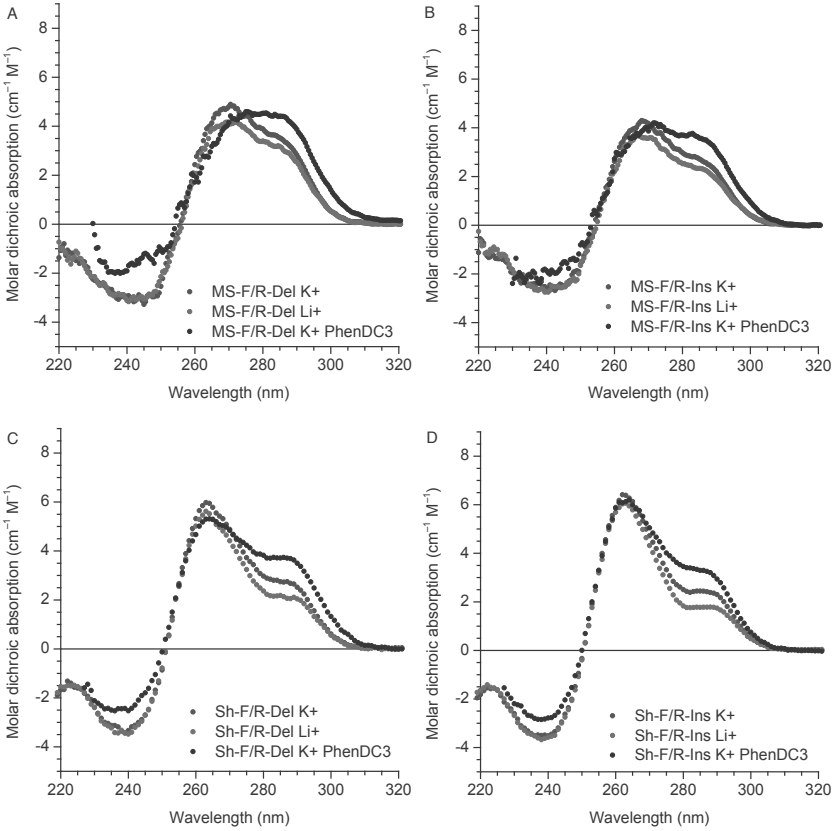


Supplementary Figure 11. Molar dichroic absorption spectra of single-stranded insertion and deletion alleles. Spectra recorded at 20 °C in 10 mM Li cacodylate buffer

(pH 7.2) supplemented with 100 mM LiCl (Li^+ conditions, **red curves**), 100 mM KCl (K^+ conditions, **blue curves**), or 100 mM KCl + 10 μM PhenDC3 (**black curves**).

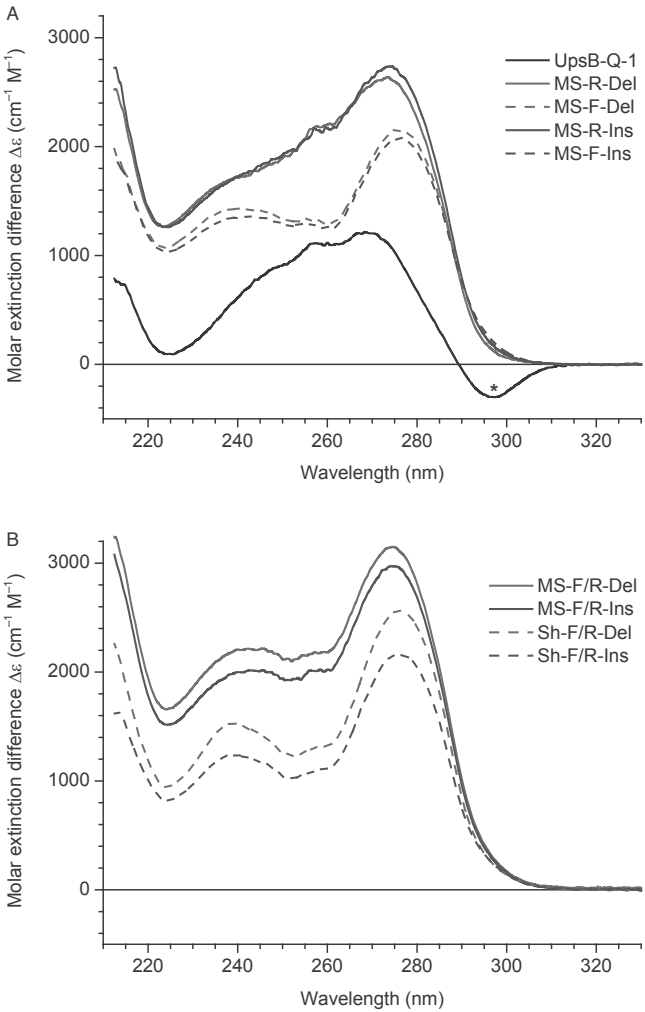
A-B Reverse (transcribed, **a**) and forward (**b**) strands of deletion (Del) allele.

C-D The same for insertion (Ins) allele. Oligonucleotide sequences are provided in **Supplementary Table 14**; Deletion: risk, Insertion: protective allele. (**e**) A 34-mer oligonucleotide UpsB-Q--CAGGGTTAAGGGTATAACTTTAGGGGTTAGGGTT-[19] was used as a positive control: the difference between the spectra observed in K^+ and Li^+ conditions is indicative of formation of a G4 structure in K^+ conditions.



Supplementary Figure 12. Molar dichroic absorption spectra of double-stranded insertion and deletion alleles.

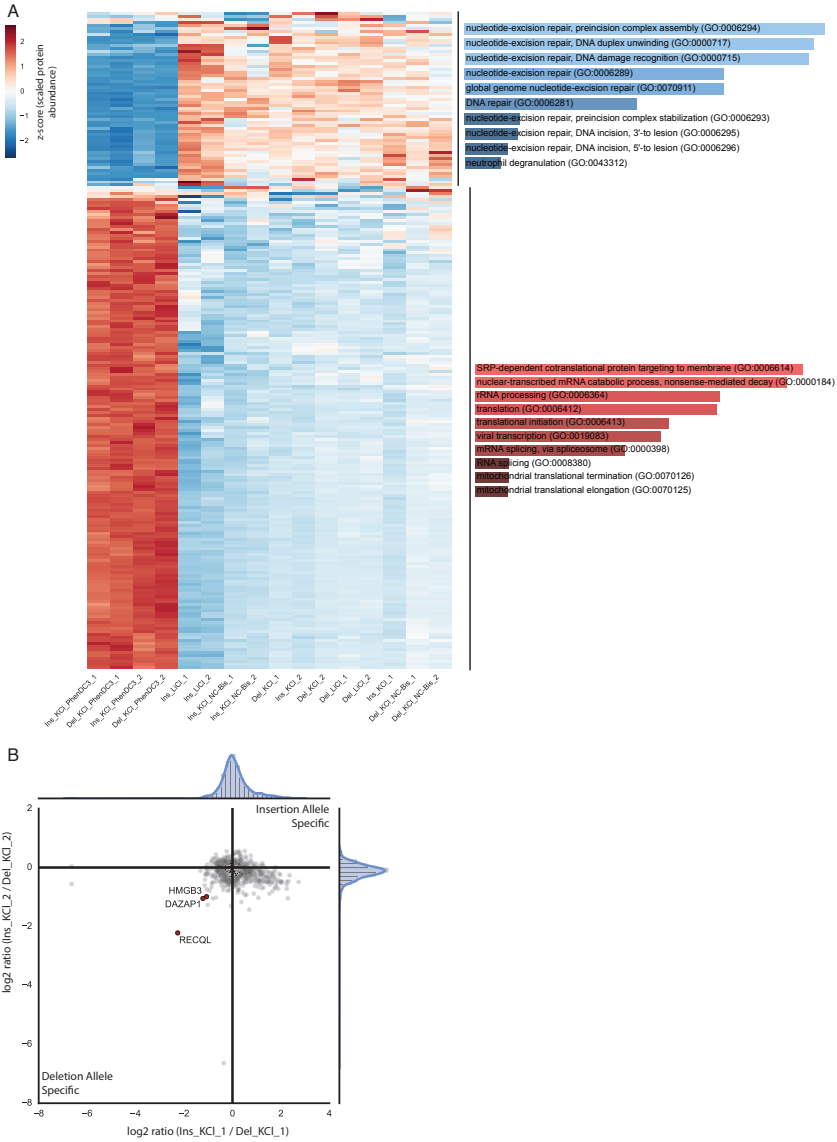
Experimental conditions and plot designations as for **Supplementary Figure 9** above. (**a–b**) Double-stranded probes, as used for mass-spec, corresponding to deletion (Del, **a**) and insertion (Ins, **b**) alleles. (**c–d**) Truncated probes corresponding to deletion (Del, **c**) and insertion (Ins, **d**) alleles. Probe sequences are listed in **Supplementary Table 14.**



Supplementary Figure 13. Molar thermal difference spectra $\Delta\varepsilon TDS = \Delta\varepsilon 95C - \Delta\varepsilon 20C$) of insertion and deletion alleles.

A **Single-stranded oligonucleotides.** MS-R- Del or MS-R -Ins: reverse (transcribed) strands, MS-F-Del or MS-F-Ins: forward strands of deletion (Del) or insertion (Ins) alleles as used for mass-spectrometry. Oligonucleotide UpsB-Q-1 (sequence: cf. **Supplementary Figure 9** above) was used as a positive control: negative peak labeled with an asterisk gives evidence of formation of a G4 structure.

B **Double-stranded (F/R) oligonucleotides** formed by annealing of reverse and forward probes used above (MS-F/R-Del, MS-F/R-Ins) or truncated sequences (Sh-F/R-Del, Sh-F/R-Ins). Oligonucleotide sequences are provided in **Supplementary Table 14**. Deletion: risk, Insertion: protective allele.



Supplementary Figure 14. Chemical perturbation of PARP1 indel DNA secondary structure recruits and repels binding proteins with little allelic preference.

Quantitative isobaric TMT labeling was used to identify preferential protein binders of the PARP1 rs144361550 ssDNA insertion or deletion allele after chemical perturbation of possible DNA secondary structures. LiCl was used to distupt potential G-quadruplex structure, while PhenDC3 was used to stabilize G-quadruplex like structures. NC-Bis was used as a negative control compound.

A Heatmap of protein binding (based on a readout of row Z-score normalized normalized TMT reporter ion abundance) for rs144361550 insertion and deletion baits in different structure perturbing binding conditions. Two main clusters were identified, distiguished as either repelled (top cluster) or recruited (bottom cluster) by PhenDC3 ligand in a non-allele specific fashion. GO term enrichment per cluster is indicated on the right of the heatmap.

B Two-dimensional interaction plot of data from (A) for the rs144361550 ssDNA insertion v deletion allele. For rs144361550 ssDNA, in contrast to dsDNA, RECQL prefers the deletion allele in KCl protein binding buffer conditions (minus PhenDC3 or NC-Bis). Outliers are called as before, using 1.5 IQRs in both replicates as an outlier calling criterion.

Supplementary Table 1 Chr1q42.1 locus eQTL genes for rs3219090 in UACC melanoma cell lines

Gene (probe)[a]	*P*-Value	Effect Size fore ach copy of risk allele, G	Standard Error
PARP1 (208644_at)	**0.0014**	**0.3653**	**0.1091**
PYCR2 (231715_s_at)	**0.0122**	**0.2567**	**0.0992**
PYCR2 (224855_at)	**0.0296**	**0.1997**	**0.0895**
EPHX1 (228549_at)	**0.0304**	**0.2761**	**0.1240**
TMEM63A (202699_s_at)	**0.0428**	**0.1811**	**0.0873**
TMEM63A (215583_at)	0.0725	0.0975	0.0532
ACBD3 (202323_s_at)	0.1153	0.2055	0.1285
SRP9 (201273_s_at)	0.1454	0.1187	0.0804
TMEM63A (202700_s_at)	0.1482	0.0788	0.0537
ADCK3 (218168_s_at)	0.2477	0.1213	0.1039
C1orf55 (1553338_at)	0.2689	0.1309	0.1171
PSEN2 (204262_s_at)	0.2928	0.1479	0.1393
LBR (201795_at)	0.3122	0.1369	0.1342
PSEN2 (211373_s_at)	0.3624	0.1467	0.1597
CDC42BPA (203794_at)	0.4880	0.0980	0.1404
CDC42BPA (214464_at)	0.5098	0.1372	0.2067
ENAH (222434_at)	0.5539	-0.0939	0.1574
ACBD3 (202324_s_at)	0.5731	0.0615	0.1084
ENAH (217820_s_at)	0.6924	0.0729	0.1833
ENAH (222433_at)	0.7180	0.0594	0.1636
EPHX1 (202017_at)	0.7602	0.0648	0.2112
ENAH (228310_at)	0.9282	0.0184	0.2033
ITPKB (203723_at)	0.9994	0.0002	0.3233

a – Gene name followed by Affymetrix U133Plus2 expression microarray probe ID. *PARP1* and other nominally significant ($P < 0.05$) eQTL genes are in bold. Bonferroni-corrected *P*-value threshold for testing 14 genes is $P < 3.6 \times 10^{-3}$

Supplementary Table 2. Chr1q42.1 locus eQTL genes for rs3219090 in TCGA melanomas

Gene[a]	*P*-Value[b]	Effect Size (risk allele, G)
PARP1	**0.00392**	**393.63**
ACBD3	0.35784	-39.02
CDC42BPA	0.35986	59.93
LIN9	0.40174	-4.89
LBR	0.41458	33.70
ENAH	0.44541	-96.02
ITPKB	0.53837	294.57
SDE2	0.63349	-9.43
DNAH14	0.63800	2.23
PSEN2	0.64290	-48.24
SRP9	0.72069	-25.43
TMEM63A	**0.77179**	**11.79**
LEFTY2	0.80346	1.27
ADCK3	0.81551	8.64
EPHX1	**0.90889**	**56.79**
H3F3A	0.97950	2.65

a – *PARP1* and other nominally significant ($P < 0.05$) eQTL genes from the discovery set (UACC melanoma cell lines) are in bold.

b – Linear regression with gene-specific regional copy number as a covariate

Supplementary Table 3. Chr1q42.1 locus eQTL genes for rs3219090 in GTEx skin tissues (Sun-exposed)

Gene[a]	*P*-Value	Effect Size (risk allele, G)
PARP1	**0.0002**	**0.1300**
PYCR2	**0.0170**	**-0.1000**
DNAH14	0.0260	-0.1400
PSEN2	0.0770	0.0870
SDE2	0.3200	-0.0370
CDC42BPA	0.4500	-0.0400
LBR	0.4500	-0.0370
H3F3A	0.5800	0.0410
H3F3AP4	0.6000	0.0170
ADCK3	0.6300	-0.0200
LEFTY2	0.6500	-0.0300
LIN9	0.7600	-0.0130
ENAH	0.8100	-0.0085
ITPKB	0.8100	-0.0120
C1orf95	0.8300	0.0130
TMEM63A	0.8700	0.0050
ITPKB-IT1	0.9500	-0.0052
ACBD3	0.9700	0.0015
SRP9	0.9900	-0.0004
EPHX1	**1.0000**	**0.0003**

a – *PARP1* and other nominally significant ($P < 0.05$) eQTL genes from the discovery set (UACC melanoma cell lines) are in bold.

Supplementary Table 4. Chr1q42.1 locus eQTL genes for rs3219090 in GTEx skin tissues (NOT Sun-exposed)

Gene[a]	*P*-Value	Effect Size (risk allele, G)
PARP1	**0.0110**	**0.1200**
PYCR2	**0.0120**	**-0.1900**
LEFTY2	0.0410	0.2100
PSEN2	0.0990	0.1100
LIN9	0.1300	0.0940
TMEM63A	**0.1500**	**-0.0610**
H3F3A	0.1600	0.1300
ADCK3	0.1700	-0.0670
SRP9	0.2000	0.0480
DNAH14	0.3500	-0.0750
ACBD3	0.3600	-0.0430
ENAH	0.4700	0.0360
ITPKB-IT1	0.4900	-0.0790
EPHX1	**0.5000**	**0.0300**
CDC42BPA	0.5700	-0.0480
C1orf95	0.6000	0.0390
LBR	0.6800	-0.0280
H3F3AP4	0.7600	0.0190
ITPKB	0.8300	0.0140
SDE2	0.9100	0.0051

a – *PARP1* and other nominally significant ($P < 0.05$) eQTL genes from the discovery set (UACC melanoma cell lines) are in bold.

Supplementary Table 5. List of 65 variants highly linked with the GWAS lead SNPs

SNP ID	Location	Allele 1	Allele 2	Max. r^{2a}	P-value [b], meta-analysis	Melanocyte histone mark [c] (n=3)	Melanoma histone mark [d] (n=11)
rs144361550	Chr1:226595403	AGGGCCC	A	1.00	4.85E-11	100%	100%
rs1417765	Chr1:226591243	G	T	1.00	4.73E-12	100%	45%
rs2695240	Chr1:226607336	C	T	1.00	3.28E-13	67%	27%
rs2570370	Chr1:226607385	C	T	1.00	3.48E-13	67%	27%
rs1341336	Chr1:226596389	G	A	1.00	8.32E-13	100%	100%
rs2793657	Chr1:226590972	C	T	1.00	1.53E-12	100%	91%
rs1858548	Chr1:226607774	G	A	1.00	3.56E-13	67%	36%
rs35380305	Chr1:226607796	G	GC	1.00	2.25E-10	67%	18%
rs1858549	Chr1:226607892	A	G	1.00	3.59E-13	67%	27%
rs2666428	Chr1:226589709	T	C	1.00	5.27E-12	100%	18%
rs2048426	Chr1:226594301	T	C	1.00	7.82E-13	100%	55%
rs2136875	Chr1:226606536	A	G	1.00	3.18E-13	67%	9%
rs878366	Chr1:226537868	C	T	0.63		67%	36%
rs577289790	Chr1:226607954	CTT	C	0.83		67%	27%
rs35242305	Chr1:226604236	C	CT	1.00		0%	0%
rs2793380	Chr1:226588977	G	C	1.00	6.39E-12		
rs2377313	Chr1:226537368	T	A	0.95	2.13E-11		
rs2570369	Chr1:226605765	C	T	1.00	1.26E-12		
rs1858550	Chr1:226608104	A	C	1.00	1.68E-13		
rs201078005	Chr1:226608615	C	CTTTTTT	0.81			
rs6677172	Chr1:226542114	G	C	0.64			
rs10915986	Chr1:226543113	C	T	0.64			
rs12567614	Chr1:226544420	T	C	0.64			
rs11801168	Chr1:226544831	C	T	0.64			
rs747657	Chr1:226550924	G	C	1.00	5.54E-12		
rs3219119	Chr1:226556443	A	T	1.00	5.47E-12		
rs3219112	Chr1:226557504	A	C	1.00	5.64E-12		
rs3835704	Chr1:226559572	TA	T	1.00	1.95E-10		

Supplementary Table 5. List of 65 variants highly linked with the GWAS lead SNPs *(Continued)*

SNP ID	Location	Allele 1	Allele 2	Max. r^{2a}	P-value [b], meta-analysis	Melanocyte histone mark [c] (n=3)	Melanoma histone mark [d] (n=11)
rs3754376	Chr1:226561064	C	A	1.00	5.00E-12		
rs3219090	Chr1:226564691	T	C	1.00	7.10E-12		
rs1805414	Chr1:226573364	G	A	0.98	5.48E-12		
rs2027439	Chr1:226575749	C	T	0.98	5.27E-12		
rs2048424	Chr1:226583007	C	G	1.00	5.17E-12		
rs2793382	Chr1:226583417	T	C	1.00	4.32E-12		
rs10631977	Chr1:226585165	CAAG	C	1.00			
rs59275599	Chr1:226586228	T	TAA	1.00			
rs11366269	Chr1:226596936	A	AT	0.96			
rs2793378	Chr1:226597341	G	A	0.98	7.80E-13		
rs1104893	Chr1:226598652	G	A	0.98	1.12E-12		
rs2793377	Chr1:226600639	C	T	0.98	1.13E-12		
rs35170928	Chr1:226600660	ACT	A	0.98	1.32E-10		
rs2377013	Chr1:226600665	A	C	0.98			
rs11304925	Chr1:226601001	C	CT	0.89			
rs2249844	Chr1:226601135	C	T	0.98	1.89E-12		
rs1397550	Chr1:226601242	C	T	0.98	1.26E-12		
rs1828446	Chr1:226601320	A	G	0.98	1.13E-12		
rs1828445	Chr1:226601401	G	A	0.98	1.12E-12		
rs2793654	Chr1:226601706	A	T	0.98	1.82E-12		
rs2666427	Chr1:226601874	A	G	0.98	1.11E-12		
rs1865222	Chr1:226601884	C	A	0.96	1.10E-12		
rs1865221	Chr1:226601974	T	G	0.98	1.78E-12		
rs1865220	Chr1:226602286	C	T	0.98	1.74E-12		
rs2695235	Chr1:226602556	G	A	0.98	1.17E-12		

a - Max r^2 is the highest r^2 value for each SNP among (1) r^2 with GWAS lead SNP rs3219090 in 1KG EUR, (2) 1KG CEU, (3) r^2 with meta-analysis [4] lead SNP rs1858550 in 1KG EUR, or (4)1KG CEU.
b - *P*-value for melanoma association from meta- analysis[4]. Blank: *P*-value is higher than 1×10^{-8}
c - Percentage of individual melanocyte samples with evidence of histone marks (H3K4me1, H3K4me3, or H3K27ac) for the region encompassing each SNP in ENCODE/Roadmap data[5] .
d - Percentage of melanoma short-term cultures or cell lines with evidence of histone mark (H3K27ac) for the region encompassing each SNP[8]

Supplementary Table 6. Nomination of *PARP1* functional risk variants

SNP ID	Max[a] r^2	P-value (meta-analysis)[b]	Melanocyte DHS (n=3)[c]	Melanoma DHS (n=13)[d]	Assignment	Candidate for validation
rs144361550	1	4.85E-11	100%	62%	**Strong supported Open Chromatin**	Yes
rs1417765	1	4.73E-12	100%	69%	**Strong supported Open Chromatin**	Yes
rs2695240	1	3.28E-13	100%	62%	**Strong supported Open Chromatin**	Yes
rs2570370	1	3.48E-13	100%	69%	**Strong supported Open Chromatin**	Yes
rs1341336	1	8.32E-13	67%	23%	Weak supported Open Chromatin	
rs2793657	1	1.53E-12	67%	23%	Weak supported Open Chromatin	
rs1858548	1	3.56E-13	67%	0%	Weak supported Open Chromatin	
rs35380305	1	2.25E-10	67%	0%	Weak supported Open Chromatin	
rs1858549	1	3.59E-13	67%	0%	Weak supported Open Chromatin	
rs2666428	1	5.27E-12	33%	0%	Weak supported Open Chromatin	
rs2048426	1	7.82E-13	33%	31%	Weak supported Open Chromatin	
rs2136875	1	3.18E-13	33%	8%	Weak supported Open Chromatin	
rs878366	0.634		33%	8%	Weak supported Open Chromatin	
rs577289790	0.826		33%	0%	Weak supported Open Chromatin	
rs35242305	0.996		33%	8%	Weak supported Open Chromatin	

a - Max r^2 is the highest r^2 value for each SNP among (1) r^2 with GWAS lead SNP rs3219090 in 1KG EUR, (2) 1KG CEU, (3) r^2 with meta-analysis [4] lead SNP rs1858550 in 1KG EUR, or (4)1KG CEU.
b - *P*-value for melanoma association from meta- analysis[4]. Blank: *P*-value is higher than 1 x 10^{-8}
c - Percentage of individual melanocyte samples with evidence of open chromatin for the region encompassing each SNP in ENCODE/Roadmap data[5]
d -Percentage of melanoma short-term cultures or cell lines with evidence of open chromatin for the region encompassing each SNP[5,8]

Supplementary Table 7. Summary of experimental validation for *PARP1* gene-regulatory candidate functional variants

	rs1417765	rs144361550	rs2695240	rs2570370
EMSA (p) Specific binding[a]	Protective allele	Protective allele	Risk allele	Not specific
EMSA (m) Specific binding[b]	Protective allele	Protective allele	Risk allele	Not specific
Luc (p) Higher activity[c]	Protective allele	No difference	Risk allele	NA
Luc (m) Higher activity[d]	Protective allele	Risk allele	Risk allele	NA
Transcriptional activity (p)[e]	Weak	Weak	Weak	NA
Transcriptional activity (m)[f]	Weak	**Strong**	Negative	NA
Direction (allele)[g]	Not consistent	**Consistent**	**Consistent**	NA

a - alleles that display specific or favorable binding to nuclear proteins from primary melanocytes
b - alleles that display specific or favorable binding to nuclear proteins from melanoma cell lines
c - alleles that display significantly higher luciferase activity when transfected to primary melanocytes
d - alleles that display significantly higher luciferase activity when transfected to melanoma cell lines
e - luciferase activity of the DHS region encompassing each SNP in primary melanocytes compared to the minimal promoter control
f - luciferase activity of the DHS region encompassing each SNP in melanoma cell lines compared to the minimal promoter control
g - denotes whether if the risk allele displays higher luciferase activity consistent with expression data (eQTL and allelic imbalance)

Supplementary Table 9. Validation of rs144361550 genotype and re-assessment of LD with rs3219090

r^2 with rs3219090[a]	rs1417765	rs144361550	rs2695240	rs2570370
1KG, phase1 EUR (n=379)	1.00	**0.667**	0.905	0.905
1KG, phase1 CEU (n=85)	1.00	**0.770**	0.869	0.869
1KG, phase3 EUR (n=503)	1.00	**0.947**	0.910	0.910
1KG, phase3 CEU (n=99)	1.00	**1.00**	0.873	0.873
DCEG reference set (healthy, EUR, n=745)[c]		**0.94** [b]		
Hapmap CEU (58 founders)[d]		**1.00** [b]		

a - r^2 values were calculated using PLINK with 1000 genomes (1KG) genotypes obtained from 1KG, phase1 and phase3 databases.
b - rs144361550 was genotyped for 30 Hapmap CEU trios and a 745 sample imputation reference panel European decent [9] using fragment analysis on CE with two different dyes (6-FAM and VIC) and two different fragment sizes.
c - rs3219090 genotypes for 745 reference individuals were determined by direct genotyping using Taqman genotyping assay.
d - rs3219090 genotypes for 58 founders were obtained from Hapmap Rel28 PhaseII+III, Aug10, b36

Supplementary Table 10. Insertion-specific binding protein identified by mass spectrometry

Protein names	Mean Ins/Del Ratio in Mass-spec[a]	Putative G4-related function	Reference	Antibody super-shift	EMSA with purified protein
CIRBP	28.1			No super-shift	No allelic binding
NCL	20.6	G4 binding (stabilizing)	Gonzalez et al[10]	No super-shift	Allelic binding
HNRNPD	20.3	G4 binding	Dempsey et al[11]	No super-shift	No allelic binding
SRSF3	19.7			No super-shift	
SRSF7	16.9				
SUB1	12.8				
RBM39	10.9				
PDCD11	9.6				
PCBP1	9.5				
RBM14	7.7				
ZC3HAV1	7.7				
RPA1	7.6	G4 unwinding helicase	Safa et al[12]	No super-shift	No allelic binding
TOP3A	7.6	G4 unwinding helicase	Temime-Smaali et al[13]	No super-shift	
DHX36	6.7	G4 unwinding helicase	Chen et al[14]	No super-shift	
RPA3	6.2	G4 unwinding helicase	Safa et al[12]	No super-shift	
RFC3	6.4				
RECQL	4.7	G4 unwinding helicase	Huber et al[15]	Super-shift	Allelic binding

a - mean ratio of heavy/light-labeled peptides from two experiments by label swapping

Supplementary Table 11. G-quadruplex forming potential prediction

A) Intramolecular G4 structures

Query Name	Deletion allele	Insertion allele
Query Sequence[a]	GAGCGAGCGGGCCCGGGCCC TCGGAGCGGCACTTGGGGCC	GAGCGAGCGGGCCCGGGCCC GGGCCCTCGGAGCGCACTT GGGGCC
Length (number of nucleotide)	40	46
QGRS found[b]	0	1
G-Score[c]	NA	58

QGRS Mapper (http://bioinformatics.ramapo.edu/QGRS/analyze.php)[20] was used with the following search parameters: QGRS Max Length: 45, Min G-Group Size: 3, Loop size: from 0 to 36.
a - Query sequence is the same as the probe sequences used for mass-spectrometry. Sequence from the reverse (transcribed) strand was used for both alleles. Predicted G-quadruplex strings are underlined and highlighted.
b - Number of non-overlapping QGRS (putative Quadruplex forming G-Rich Sequences) found in the query sequence
c - Likelihood score for forming a stable G-quadruplex. NA: score not available

B) Intermolecular G4 structures

Query[a]	Sequence (5ǫ–3ǫ)[b]	Putative G4 topology[c]
Del allele	(F) GGCCCCAAGTGCCGCTCCGA**GGGCCCGGG**CCCGCTCGCTC	ABAB
	(R) GAGCGAG**CGGGCCCGGGCCCT**CGGAGCGGCACTTGGGGCC	
Ins allele	(F) GGCCCCAAGTGCCGCTCCGA**GGGCCCGGGCCCGGG**CCCGCTCGCTC	ABAB, ABBB, or AAAB
	(R)GAGCGAG**CGGGCCCGGGCCCGGGCCCT**CGGAGCGGCACTTGGGGCC	

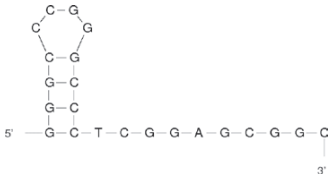
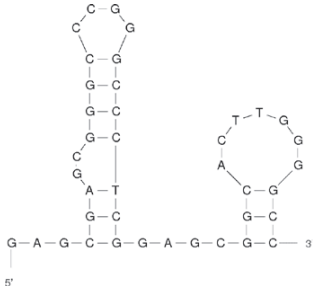
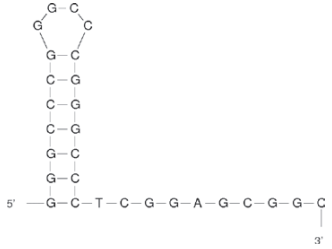
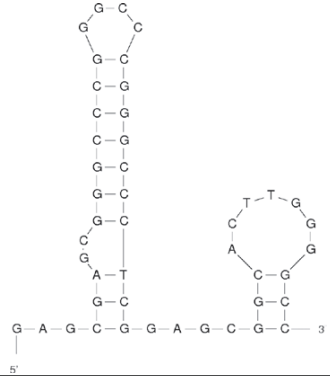
Sequences were analyzed according to algorithms described by Kudlicki[21].
a - Query is the double-stranded sequence corresponding to probes described above. G runs which can participate in formation of G-quartets are underlined and highlighted.
b - Truncated sequences used for a biophysical study are typed in bold (cf. **Supplementary Table 14** below).
c - Topological description of putative G4 structures using notation from Kudlicki[21].

Supplementary Table 12. *PARP1* and *MITF* target gene expression correlation in melanoma cell lines

Gene Symbol[a]	Probe ID[b]	vs *MITF* Pearson *r* (n=59)	*P*-value[c]	vs *PARP1* Pearson *r* (n=59)	*P*-value[c]	vs *PARP1* Pearson *r* (n=23, *MITF*-high)	*P*-value[c]
CDK2	211804_s_at	0.85	1.67E-17	0.25	0.0588	0.37	0.0827
TBX2	40560_at	0.70	5.09E-10	0.23	0.0779	0.32	0.1420
RAB27A	209514_s_at	0.69	1.60E-09	0.19	0.1487	0.15	0.4940
EDNRB	204271_s_at	0.68	1.40E-08	0.11	0.4436	0.43	0.0394
MC1R	205458_at	0.42	1.04E-03	-0.01	0.9298	-0.07	0.7443

a – *MITF* target genes were selected based on 13 melanocyte-specific target genes that were validated by Hoek et al[22]. Five of those thirteen genes are expressed above background levels in our 59 early-passage melanoma cell lines and used for testing correlation with *PARP1* levels.
b – probe ID from Affymetrix U133Plus2 expression microarray
c – Correlation analyses were performed regardless of genomic copy number.

Supplementary Table 13. Mfold[a]-predicted lowest-energy secondary structures of oligonucleotides used for G4 analyses.

Oligonucleotide	Structure	ΔG / kcal mol^{-1}
EMSA-R-Del		−3.94
MS-R-Del		−6.12
EMSA-R-Ins		−9.79
MS-R-Ins		−11.97

a - http://unafold.rna.albany.edu/?q=mfold [23]

Supplementary Table 14. Oligonucleotide sequences

rs144361550 Oligo probes	Forward (5'-3')	Reverse (5'-3')
Deletion EMSA and luciferase construct (22bp)	GCCGCTCCGAGGGCCCGGG CCC	GGGCCCGGGCCCTCGGAGCGGC
Insertion EMSA and luciferase construct (28bp)	GCCGCTCCGAGGGCCCGGG CCCGGGCCC	GGGCCCGGGCCCGGGCCCTC GGAGCGGC
Deletion mass spec, CD and TDS (40bp)	GGCCCCAAGTGCCGCTCCG AGGGCCCGGGCCCGCTC GCTC	GAGCGAGCGGGCCCGGG CCCTCGGAGCGGCACTT GGGGCC
Insertion mass spec, CD and TDS (46bp)	GGCCCCAAGTGCCGCTC CGAGG GCCCGGGCCCGGG CCCGCTCGCTC	GAGCGAGCGGGCCCGGG CCCGGGCCCTCGGAGCGG CACTTGGGGCC
Truncated deletion CD and TDS (14bp)	AGGGCCCGGGCCCG	CGGGCCCGGGCCCT
Truncated insertion CD and TDS (20 bp)	AGGGCCCGGGCCCGGG CCCG	CGGGCCCGGGCCCGGGCCCT
PCR primers for re-genotyping rs144361550	**Forward (5'-3')**	**Reverse (5'-3')**
156bp amplicon (6-FAM)	GCAACATCAGCAAAACC TTC	CCCGGGTTAACTGTGTCC
240bp amplicon (VIC)	CCACCCAGAAAGGAGAA GAG	GTTAACTGTGTCCGGGAAGG
PCR primers for rs144361550 ChIP qPCR	**Forward (5'-3')**	**Reverse (5'-3')**
RECQL positive control (Pos)	CTCCACCCCCAAGGAAAA AG	GGCAGGGTCCCATGCA
rs144361550 locus (Indel)	GAAGAGGCTCCTCGTTTT CAC	TTCCGGAAGGTTTTGCTG
EMSA probes	**Forward (5'-3')**	**Reverse (5'-3')**
rs1417765 (21bp)	ATGTAGTGTG[T/G] GTCCCTGCTC	GAGCAGGGAC[A/C] CACACTACAT
rs2695240 (21bp)	GGGGCTGATG[T/C] GGGAGCCTCG	CGAGGCTCCC[A/G] CATCAGCCCC
rs2570370 (21bp)	GACAGCAAAG[T/C] CATGAGAAAT	ATTTCTCATG[A/G] CTTTGCTGTC
primers for rs2695240-rs2570370 haplotype probe (186bp amplicon)	CAGTCATTAAGAAAA CAAGGACAAAG	ATGAGACACCCTGG AAATAAATG
PCR primers for luciferase constructs	**Forward (5'-3')**	**Reverse (5'-3')**
rs144361550 (905bp)	CTGGGACAGAAC AATCAAAGG	GTCGCCACCATCCATGTAG
rs1417765 (557bp)	TTCCAGAGTATTTT CCCTGTCC	TTAGTAGCAATGGGGCTTCAC
rs2695240-rs2570370 (711bp)	TCAGGAAATCTGC ACTCTGG	GTCAGGGCAAGTCCATTAGG
***MITF*-M promoter ChIP primers**	**Forward (5'-3')**	**Reverse (5'-3')**
Primer1	AACATGAATCTCTTTTC TTTTTAAGTG	CAATCTCATATTGTTTC AAATGACTG
Primer2	TGCATTATCCTGGG CATTTAG	TTGTGAACAAACAG ATATAGTTTCCAG

Supplementary Table 14. Oligonucleotide sequences *(Continued)*

***MITF*-M promoter ChIP primers**	**Forward (5'-3')**	**Reverse (5'-3')**
Primer3	ATGCTTTGTACAGTGTTT AGCAC	AAAGCATGAGCATTTTTGCTGAG
Primer4	CCACTTCTGTGTGTGCT ATGTTC	TCCGACTGCAGGATGACTATC
Primer5	GGGGCATTCTGCTATTAACC	ATTTTTCCCCCTGGCTTG
Primer6	AAAAGGCCCTTATGTGAA CG	GGTAGACTATCCCTCCCTCTAC
Primer7	TGGTGTCTCGGGATACCTTG	TGAGTCAGAATAAATCTCA CCTGATAG
Primer8	TGCTCTTTTAATGCTGTTT ATTATTTG	TGAGCAATGAACAGGAGCTG
Primer9	CATCAGCTCCTGTTCATTGC	TTTCAAATGCATAACAC TTACACG
Primer10	CAGGGAAATAAAATAGGG CAAAG	TTTCAGACGGCTCTTCCTTC
Primer11	GGGCAGGCAGTTTAGCA TAG	CATTGGGTTCGAGGATTTTC
Primer12	GGGCTATA-AGCTTTTTCAACTGG	CGTGGGGGATACCTAGTGAG
Bisulfite sequencing primers	**Forward (5'-3')**	**Reverse (5'-3')**
Cg06640206 (F3 and R3)	TTTTTTAAAGGGGTATTTT GTTATT AATTTATTGTTG	AAACACCACCRAAAACTTT ATCAC AAAAACCCTAC
Cg11038507 (F7 and R7)	ATGTAGTTAAGAATAAGGT GTATA TTAAGATTAGGATG	AAATAACRAACTATCAAAA TCAAACTCACTATC
Primers for *MITF*-M promoter luciferase construct	**Forward (5'-3') with KpnI overhang (lower case)**	**Reverse (5'-3')**
MITF.674 construct	actgaactggtaccATTCTCAGC AAAAATGCTCATGC	CCAAGCTTACTTAGATCTCGAG

References

1 Nalla, V.K. & Rogan, P.K. Automated splicing mutation analysis by information theory. *Hum Mutat* **25**, 334-42 (2005).

2 Rogan, P.K., Faux, B.M. & Schneider, T.D. Information analysis of human splice site mutations. *Hum Mutat* **12**, 153-71 (1998).

3 Barrett, J.H. *et al.* Fine mapping of genetic susceptibility loci for melanoma reveals a mixture of single variant and multiple variant regions. *Int J Cancer* **136**, 1351-60 (2015).

4 Law, M.H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat Genet* **47**, 987-95 (2015).

5 Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

6 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

7 Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

8 Verfaillie, A. *et al.* Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat Commun* **6**, 6683 (2015).

9 Wang, Z. *et al.* Improved imputation of common and uncommon SNPs with a new reference set. *Nat Genet* **44**, 6-7 (2012).

10 Gonzalez, V., Guo, K., Hurley, L. & Sun, D. Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. *J Biol Chem* **284**, 23622-35 (2009).

11 Dempsey, L.A., Sun, H., Hanakahi, L.A. & Maizels, N. G4 DNA binding by LR1 and its subunits, nucleolin and hnRNP D, A role for G-G pairing in immunoglobulin switch recombination. *J Biol Chem* **274**, 1066-71 (1999).

12 Safa, L. *et al.* Binding polarity of RPA to telomeric sequences and influence of G-quadruplex stability. *Biochimie* **103**, 80-8 (2014).

13 Temime-Smaali, N. *et al.* The G-quadruplex ligand telomestatin impairs binding of topoisomerase IIIalpha to G-quadruplex-forming oligonucleotides and uncaps telomeres in ALT cells. *PLoS One* **4**, e6919 (2009).

14 Chen, M.C., Murat, P., Abecassis, K., Ferre-D'Amare, A.R. & Balasubramanian, S. Insights into the mechanism of a G-quadruplex-unwinding DEAH-box helicase. *Nucleic Acids Res* **43**, 2223-31 (2015).

15 Huber, M.D., Duquette, M.L., Shiels, J.C. & Maizels, N. A conserved G4 DNA binding domain in RecQ family helicases. *J Mol Biol* **358**, 1071-80 (2006).

16 De Cian, A., Delemos, E., Mergny, J.L., Teulade-Fichou, M.P. & Monchaud, D. Highly efficient G-quadruplex recognition by bisquinolinium compounds. *J Am Chem Soc* **129**, 1856-7 (2007).

17 Watanabe, A., Takeda, K., Ploplis, B. & Tachibana, M. Epistatic relationship between Waardenburg syndrome genes MITF and PAX3. *Nat Genet* **18**, 283-6 (1998).

18 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).

19 Smargiasso, N. *et al.* Putative DNA G-quadruplex formation within the promoters of Plasmodium falciparum var genes. *BMC Genomics* **10**, 362 (2009).

20 Kikin, O., D'Antonio, L. & Bagga, P.S. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* **34**, W676-82 (2006).

21 Kudlicki, A.S. G-Quadruplexes Involving Both Strands of Genomic DNA Are Highly Abundant and Colocalize with Functional Sites in the Human Genome. *PLoS One* **11**, e0146174 (2016).

22 Hoek, K.S. *et al.* Novel MITF targets identified using a two-step DNA microarray strategy. *Pigment Cell Melanoma Res* **21**, 665-76 (2008).

23 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-15 (2003).

Chapter 4

Global profiling of protein-DNA and protein-nucleosome binding affinities using quantitative mass spectrometry

Modified from:
Global profiling of protein-DNA and protein-nucleosome binding affinities using quantitative mass spectrometry.

Matthew M Makowski, Cathrin Gräwe[#], Benjamin M Foster[#], Nhuong V Nguyen[#], Till Bartke[*], Michiel Vermeulen[*]

Nature Communications. 2018.

Abstract

Interaction proteomics studies have provided fundamental insights into multimeric biomolecular assemblies and cell-scale molecular networks. Significant recent developments in mass spectrometry-based interaction proteomics have been fueled by rapid advances in label-free, isotopic, and isobaric quantitation workflows. Here, we report a quantitative protein-DNA and protein-nucleosome binding assay that uses affinity purifications from nuclear extracts coupled with isobaric chemical labeling and mass spectrometry to quantify apparent binding affinities proteome-wide. We use this assay with a variety of DNA and nucleosome baits to quantify apparent binding affinities of monomeric and multimeric transcription factors and chromatin remodeling complexes.

Introduction

Interaction proteomics, via mass spectrometry, has contributed invaluably to the identification of physical associations between biological molecules and the translation thereof into cellular protein networks[1-3]. Interaction proteomics methodologies are generally semi-quantitative and use label-free or chemical labeling based relative quantification to call interactions as "outliers" from a background of non-specific identifications[4,5]. Thus, interactions are regularly reported in a binary "on/off" manner, though semi-quantitative stoichiometric information is beginning to add a quantitative dimension to interaction studies. Nevertheless, a complete characterization of a functioning cell requires knowledge not only of specificity of biomolecular interactions (i.e., does an interaction occur, or not) but also of affinity (i.e., how strong, in absolute terms, is some given interaction)[6]. Typical affinity quantitation methods, such as ITC, SPR, Fluorescence Polarization, FRET, or EMSA, have been performed on a single interaction, case-by-case basis and require laborious expression and purification of recombinant proteins. Importantly, chemoproteomic approaches studying protein-small molecule interactions established the possibility of designing absolutely quantitative binding assays using semi-quantitative isobaric labeling and mass spectrometry[7,8]. Similarly, thermal proteome profiling is an innovative mass spectrometry approach that uses thermal stability shifts upon small molecule binding to estimate apparent dissociation constants proteome-wide, again with a semi-quantitative isobaric labeling strategy[9,10]. Other protein-centric studies have focused on measuring the DNA binding landscape of a single, or a few, transcription factors[11,12]. However, the inverse problem of interrogating the quantitative protein binding landscape of DNA sequences of interest has received considerably less attention and is still dominated by semi-quantitative workflows[13,14]. Here, we present a method for determining, proteome-wide, tens to hundreds of apparent dissociation constants (K_d^{app}) of nuclear proteins for DNA and nucleosome ligands simultaneously using affinity purification from nuclear lysates and isobaric 10-plex TMT labeling coupled with mass spectrometry.

Methods

Cell culture and nuclear lysate preparation

Wild-type HeLa Kyoto cells (received from Anthony Hyman and Ina Poser, Human HeLa BAC database, Dresden, Germany) were cultured in DMEM supplemented with 10% FBS and 100 U/mL penicillin and streptomycin. Cells were periodically tested for mycoplasma contamination. Nuclear lysates were isolated as described previously[14]. Briefly, cells were lysed by swelling and mechanical force in buffer A (10 mM HEPES (pH 7.9), 1.5 mM MgCl2, 10 mM KCl and 0.15% NP40). Then, nuclei were collected by centrifugation and chemically lysed in buffer C (420 mM $NaCl_2$, 20 mM Hepes (pH 7.9), 20% (v/v) glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% NP40, EDTA-free complete protease inhibitors (CPIs, Roche), and 0.5 mM DTT). Protein concentrations were assessed by Bradford assay.

Affinity purification and sample preparation

Oligonucleotides for affinity purification were ordered as custom synthesized oligos from Integrated DNA Technologies (IDT, Supplementary Table 1). DNA and nucleosome affinity purifications were performed using a filter plate based workflow described first by Hubner et al[15]. The essential protocol is described below.

First, all dsDNA oligos were annealed by heating to 95C for 10 minutes before cooling to room temperature. Each oligo was then diluted to a working stock of 3 μM, which represented the highest concentration "reference" titration point in this study. A series of nine three-fold dilutions was then prepared, resulting in a titration series of 10 oligo concentrations ranging from 0.15 nM to 3 μM. We prepared 200 uL of oligo per titration point per replicate in this fashion. Dilutions were performed in DNA binding buffer (DBB: 1 M NaCl, 10 mM Tris pH 8, 1 mM EDTA, 0.05% NP-40).

Filter plates were first prepared with 50 uL of ethanol per well (96-well filter plate, 1.2 μM pore, Millipore/Merck MSBVS1210). Wells were then washed twice with DBB. 20 uL streptavidin-sepharose bead slurry was added to each well (GE, 10 uL beads, 3 nmol binding capacity). Wells were washed twice with DBB. 150 uL of the corresponding oligo titration point was added to each well, and oligos were immobilized to the streptavidin-conjugated beads over a

1 hour incubation at 4C while shaking on a tabletop microplate shaker. Samples were then washed once with DBB and twice with protein binding buffer (PBB: 150 mM NaCl, 50 mM Tris pH 8, 0.25% NP-40, 1 mM TCEP, and CPIs). 100 μg of HeLa nuclear lysate was then diluted to 150 uL final volume in PBB and added to each well. Samples were incubated for 2 hours at 4C while shaking on a microplate shaker, then washed six times with washing buffer (150 mM NaCl, 100 mM Triethylammonium bicarbonate, TEAB).

For competition experiments using SP/KLF wild-type and mutated oligonucleotides, beads were pre-washed as described and pre-incubated in 96-well plate format with 200 nM biotinylated SP/KLF wild-type oligonucleotide diluted in DBB for 1 hour at 4C while shaking on a tabletop microplate shaker. Free (unbiotinylated) wild-type or mutated SP/KLF oligonucleotide were prepared at three-fold diluted concentrations in PBB from 0.3 nM to 6 μM as described above. 100 μg of HeLa nuclear lysate was prepared in PBB and oligonucleotides were mixed with HeLa nuclear lysates at a ratio of 1:1 in a final volume of 150 uL, with free oligonucleotides at a final concentration of 0.15 nM to 3 μM. Samples were then washed once with DBB and twice with PBB. HeLa lysates plus free oligonucleotides were added to the beads containing 200 nM immobilized SP/KLF oligonucleotide. Proteins were incubated for 2 hours at 4C while shaking while shaking on a microplate shaker, then washed six times with washing buffer. Sample preparation for mass spectrometry was continued as described below.

For the LiCl mycG4 experiment, the oligos were heated for 10 minutes at 95C and afterwards snap-cooled on ice for 3-5 minutes before they were immobilized on the beads. The NaCl in the PBB was replaced with 150 mM LiCl for the oligo titration, protein incubation, and washing steps (PBB composition: 150 mM LiCl, 50 mM Tris pH 8, 0.25% NP-40, 1 mM TCEP, and CPIs). For the PhenDC3 mycG4 experiment, 20 μM PhenDC3, final concentration, was added to the PBB for oligo titration, protein incubation, and washing steps (PBB composition: 150 mM NaCl, 50 mM Tris pH 8, 0.25% NP-40, 1 mM TCEP, CPIs, and 20 μM PhenDC3). For the nucleosome experiments, nucleosome titration and immobilization was performed in PBB instead of DBB. All other experimental conditions were the same as described above.

Sample preparation for mass spectrometry was performed by first adding 50 uL of elution buffer (20% methanol, 80 mM TEAB, 10mM TCEP). Proteins

were incubated for 30 minutes at room temperature, alkylated with 50 mM iodoacetamide, and digested with 0.25 µg trypsin overnight while shaking on a tabletop shaker at room temperature.

For formaldehye cross-linking experiments using the mycG4 ssDNA oligonucleotide, 500 µg of HeLa nuclear lysate was prepared in a volume of 600 uL borate buffered saline (50 mM boric acid pH 8.4, 150 mM NaCl, 0.25% NP-40, 1 mM DTT, and CPIs). 500 pmols of mycG4 ssDNA oligonucleotide was added to each reaction (no DNA was added to duplicate control reactions). Reactions were incubated for 90 minutes at 4C while rotating end-over-end. Formaldehyde was added to a 1% final concentration, and reactions were incubated for 10 mins at 30C. Cross-linking was quenched by adding glycine to a final concentration of 12.5 mM and incubating for 5 minutes at 30C. 20 uL of streptavidin-sepharose bead slurry was added to each reaction, and bead reactions were incubated for 30 minutes at 4C with rotation. Each reaction was washed three times with 8 M urea prepared in 50 mM ammonium bicarbonate. Each reaction was resuspended in 100 uL of 2 M urea with 50 mM ammonium bicarbonate and 10 mM DTT. Samples were incubated for 30 minutes at room temperature with shaking on a tabletop shaker. IAA was added to each reaction to 50 mM and incubated for 10 minutes in the dark while shaking. 2.5 µg trypsin was added to each sample, and reactions were incubated overnight at room temperature while shaking. Digested samples were washed three times with 8 M urea plus 50 mM ammonium bicarbonate and three times with 50 mM ammonium bicarbonate to remove digested peptides not cross-linked to DNA oligonucleotides. 100 uL of 50 mM ammonium bicarbonate was added to each sample, and cross-links were reversed by incubating at 70C for 90 minutes. The supernatant was collected and transferred to a new tube. 0.25 ug fresh trypsin was added to each sample. Samples were incubated at room temperature for 4 hours with shaking, prepared on stageTips, and labelled by dimethyl chemical labelling on stageTips as described previously[16,17].

Isobaric labeling was performed using the 10-plex tandem mass tag (TMT) system (Thermo)[18]. 0.8 mg TMT reagent for each reporter mass was resuspended in 100 uL anhydrous acetonitrile. 10 ul resuspended TMT reagent was then added to the corresponding sample. Reactions were incubated for 1 hour in the dark before quenching for 30 minutes with 100 mM Tris pH 8.0. All ten pulldowns corresponding to all ten oligonucleotide titration points labelled by

the corresponding TMT reagent were pooled into one Eppendorf tube, acidified with trifluoroacetic acid, and desalted for mass spectrometry analysis by the C18 StageTip method[19].

Preparation of modified nucleosomes

Recombinant human core histone proteins were expressed in E. coli BL21(DE3)/RIL cells from pET21b(+) (Novagen) vectors and purified by denaturing gel filtration and ion exchange chromatography essentially as described[20]. Truncated human H3.1Δ1-31T32C protein for native chemical ligation of modified histone H3 variants was expressed in BL21(DE3)/RIL cells and purified as described[21]. Native chemical ligations were carried out in 550 µl of degassed NCL buffer (200 mM KPO4, 2 mM EDTA, 6 M Guanidine HCl) containing 1 mg of modified H3.1 aa's 1-31 thioester peptide (Cambridge Peptides), 4 mg of truncated H3.1Δ1-31T32C, 12.5 mg 4-Mercaptophenylacetic acid (MPAA) and 10 mg TCEP as reducing agent at a pH of 7.5. The reactions were incubated over night at 40C and quenched by addition of 60 µl of 1 M DTT and 700 µl 0.5% acetic acid. After a centrifugation step to remove precipitates the ligation reactions were directly loaded and purified on a reversed phase chromatography column (Perkin Elmer Aquapore RP-300 C8 250x4.6 mm i.d.) using a gradient of 45-55% B (Buffer A: 0.1% TFA in water; B: 90% acetonitrile, 0.1% TFA) over 10 column volumes. Positive fractions containing ligated full-length histone H3.1 were then combined and lyophilized. Histone octamers were refolded from the purified histones and assembled into nucleosomes with biotinylated DNA via salt deposition dialysis as described[20]. Biotinylated nucleosomal DNAs containing either one (mono-nucleosomes) or two 601 nucleosome positioning sequences[22] separated by a 50 bp linker (di-nucleosomes) were prepared as described[21]. Di-nucleosomes were assembled in the presence of MMTV A competitor DNA and a slight excess of octamers as described for longer chromatin arrays to ensure saturation of the 601 repeats[23]. The reconstituted nucleosomes were then immobilized on magnetic streptavidin beads (Dynabeads MyOne Steptavidin T1) via the biotinylated DNA, washed to remove MMTV A competitor DNA and MMTV A nucleosomes (in the case of di-nucleosomes), and directly used for affinity pull down reactions as described above. Nucleosome quality control checks are shown in Supplementary Fig. 9.

Mass spectrometry analysis

Samples containing labelled peptides were eluted from StageTips with buffer B (80% acetonitrile, 0.1% formic acid), concentrated to 5 uL by SpeedVac centrifugation at room temperature, and resuspended to 12 uL in buffer A (0.1% formic acid). Samples were separated by liquid chromatography using an Easy-nLC 1000 system (Thermo). For our first SP/KLF replicate, we used a modified gradient from 7-15% buffer B over 5 minutes, from 15%-35% buffer B over 174 minutes, from 35-50% buffer B over 5 minutes, and finally from 50-95% buffer B in 1 minute followed by 5 minutes hold at 95% buffer B. For all other replicates and experiments, the gradient was the same with the exception of the 15-35% buffer B gradient extending over 214 minutes.

Mass spectrometry analysis was performed on a Thermo Fusion Tribrid instrument using the built-in Thermo synchronous precursor selection (SPS) MS3 method, with a modified nano-HPLC gradient as described above[24,25]. Briefly, full MS scans were collected in the orbitrap at 120,000 resolution in a scan range from 380-1500 m/z. We used an AGC target of 2.0e5, and a maximum injection time of 50 ms. Peaks were selected for MS2 based on selection criteria of charge state 2-7 and an intensity threshold of 5.0e3. Dynamic exclusion was enabled, with peaks excluded after 1 scan for a duration of 70 seconds in a ±10 ppm window. MS2 was conducted in top speed data dependant acquisition mode, with precursor priority given based on highest intensity. MS2 scans were performed after isolation in the quadrupole using an isolation window of 0.7 m/z units. We used CID activation at a collision energy of 35% for fragmentation. MS2 detection was performed in the ion trap with an AGC target of 1.0e4 and a maximum injection time of 50 ms. MS2 precursors in the mass range 400-1200 m/z were selected for MS3 analysis using the Thermo TMT reagent isobaric tag loss exclusion property and excluding MS2 precursor ions 18 m/z units low and 5 m/z units high. MS3 selection was conducted in top 10 data dependent acquisition mode giving the most intense ions the highest precursor priority. MS3 ions were selected with synchronous precursor selection activated for 10 precursors. MS and MS2 isolation windows were set to 2 m/z. HCD activation was used at a collision energy of 65%. Fragment ions were detected in the orbitrap with 60,000 resolution in the scan range 120-500 m/z. For MS3, we used an AGC target of 1.0e5 and a maximum injection time of 120 ms.

Mass spectrometry analysis of formaldehyde cross-linking experiments

was performed using a 2 hour gradient with chromatography and instrument settings as reported previously[17].

Computational identification and quantification of proteins

Spectral matching to peptides, grouping of peptide identifications into proteins, and isobaric label quantification were performed using Proteome Discoverer 2.1 (Thermo). We used the built-in processing workflow "PWF_Fusion_Reporter_Based_Quan_SPS_MS3_SequestHT_Percolator" and the built-in consensus workflow "CWF_Comprehensive_Enhanced Annotation_Quan_Results", both with default settings. We used the TMT 10-plex quantification method with the 131 mass set as the control channel. In our workflow, the 131 reporter mass always corresponded to the titration point with the highest bait concentration (3 μM), with each sequentially lighter reporter tag corresponding to a threefold dilution of the next highest bait concentration. For the Sequest HT search, database parameters were enzymatic digestion with trypsin allowing two missed cleavages, a minimum peptide length of 6 amino acids and a maximum peptide length of 144 amino acids. Our search was performed against the uniprot curated human proteome (downloaded December 2015). We used a precursor mass tolerance of 10 ppm and a fragment mass tolerance of 0.6 Da. Cysteine carbamidomethylation was included as a static modification (57.021 Da), while methionine oxidation (15.995 Da) and protein N-terminal acetylation (42.011 Da) were included as dynamic modifications. We included the 6-plex TMT reagent mass (229.163 Da) as a dynamic modification on lysine, histidine, serine, and threonine, as well as the peptide N-terminus. FDR filtering was performed via percolator with a strict target FDR of 0.01 and a relaxed FDR of 0.05[26]. Strict parsimony was applied for protein grouping, and unique plus razor peptides were used for quantification. Peptide quantification normalization was applied based on total peptide amount.

For peptide searching taking into account peptide phosphorylation, we considered serine, threonine, and tyrosine (STY) phosphorylation events (79.966 Da) as possible dynamic modifications. All other parameters were unchanged.

Peptide identification and quantification of dimethyl chemical labels for formaldehyde cross-linking experiments was performed using the MaxQuant software package[27] v1.6.0.1 searching against the UniProt curated human

proteome (released June 2017). Carbamidomethylation was included as a fixed modification and methionine oxidation and protein N-terminal acetylation were included as variable modifications. Requantification was selected as a quantification parameter. Normalized peptide ratios were transformed to log2, and replicates were plotted against each other in two dimensions. Outliers were called based on inter-quartile ranges using an inter-quartile range of 1.5 in each replicate as a cutoff value.

HeLa nuclear lysate absolute proteome quantification

Absolute abundances of proteins in HeLa nuclear lysate were reported previously (Supplementary Data 2)[28]. Quantification was based on the iBAQ method as described previously using MaxQuant version 1.2.2.5[27,29-31].

Fitting of binding parameters and statistical analysis

To calculate protein binding parameters, we fit a Hill-like curve of the form:

$$\Theta = \frac{1}{\left(\frac{KdApp}{[L]}\right)^{n}+1} \qquad \text{Eq. 1}$$

where θ represents the fraction of protein bound, [L] represents the concentration of bait, K_d^{App} represents the apparent dissociation constant at which half the protein is bound to a bait molecule, and n is the Hill coefficient describing the rate at which binding saturates. Θ was observed by calculating the normalized ratio of each titration point to the 131 reporter ion signal (representing a 3 μM bait concentration), implicitly assuming that for each protein that shows K_d^{App} values in the nanomolar range we would essentially saturate binding at this titration point. Then, the signal from the 131 reporter ion represented the complete bait binding population of the entire sample, so the signal from each titration point relative to the 131 reporter ion represented the "fraction bound" of the total binding population. [L] was known from the experimental design. We fit the parameters K_d^{App} and n using non-linear least squares using the mean θ of each triplicate (SP/KLF oligo experiments and nucleosome experiments) or duplicate (motif survey experiments including mycG4 experiments). Proteins were considered for further analysis only if θ was measured for each titration point of each replicate. θ values from MS3 quantification in Proteome Discoverer were normalized by min-max scaling between 0 and 1. We used

initial parameter estimates of K_d^{App}=100 and n=1. After fitting, our data was filtered first based on the goodness of fit of the Hill-like curve. We required a r-squared value of 0.95 (0.9 in nucleosome experiments) for a linear regression between the fit binding model and the measured data, as well as a predicted fraction bound of < 0.25 for the lowest titration point and > 0.75 for the highest titration point. In other words, fit binding curves should match the data well to avoid spurious fitting, and the binding curve should nearly saturate on both sides based on the assumptions in the workflow described above. For SP/KLF competition experiments, binding parameters were fit using a Hill-Like curve essentially as described above, using an r-squared cutoff of 0.95, to estimate IC_{50} values. These IC_{50} values were then converted to K_d^{App} values using the Cheng-Prusoff correction:[8]

$$K\text{dApp\_Free} = \frac{IC50}{1+\frac{200}{K\text{dApp Immobilized}}} \quad \text{Eq. 2}$$

where $K_d^{App\_Free}$ is the calculated K_d^{App} for either the wild-type or the mutated SP/KLF oligonucleotide, IC_{50} is the fit IC_{50} value from the competition experiment, and $K_d^{App\_Immobilized}$ is the K_d^{App} value calculated from the immobilized SP/KLF wild-type oligonucleotide experiment. K_d^{App} values identified in this study are listed in Supplementary Data 1. Binding curves were plotted as the mean of replicates plus and minus error bars representing the standard error of the mean calculated using bootstrapping in the python package seaborn. Clustering of protein binding profiles was performed by hierarchical agglomerative clustering, and the mean K_d^{App} value of each sample-cluster combination was plotted in a heatmap. The number of clusters was selected by manual inspection. For statistical comparisons between multiple members of a complex or complexes, we performed a two-sided *t*-test treating all measured K_d^{App} values for that complex or paralog group as sample populations as described in the main text and figure legends.

Comparison of K_d^{App} and motif score

To compare fitted K_d^{App} values with motif scores from genome-wide binding models, we first collected all vertebrate JASPAR motifs for factors both with measured K_d^{App} values and identified in control experiments with mutated SP/KLF oligonucleotide as sequence-specific. To account for intrinsic differences in length and information content of the different motifs, we calculated the

normalized motif score, defined here as the maximum motif score for the oligo sequence used divided by the patser motif significance threshold in biopython[32].

G-quadruplex enrichment in PRC2 and NuRD ChIP-Seq peaks

We compared publically available ChIP-sequencing data for SWI/SNF, PRC2 and NuRD subunits[33,34] with publically available G4-sequencing data[35] (Supplementary Table 3). All sequencing datasets were mapped to the human genome build hg38 using the UCSC genome browser liftOver tool[36]. All G4 peaks from the plus and minus strand with overlapping coordinates were combined using bedtools[37]. We used automated permutation based testing with pybedtools[38] to look for significant correlation between ChIP-seq peaks and G4-seq peaks. We randomized ChIP-seq peaks 1000 times over the genome and each time measured the peak intersection with G4-sequencing peaks. We then calculated an empirical p-value by comparing the number of true intersected peaks between SWI/SNF, PRC2, or NuRD subunits and G4-seq peaks with these 1000 randomized intersections.

Protein cloning, expression, and western blotting

DNA pulldowns were performed as described above for western blotting. For the recombinant SP3 and SP3+HeLa spike-in pulldowns, 50 ng SP3 protein/pulldown was used, and the protein binding step was performed in protein binding buffer supplemented with 10 ng/uL BSA (final concentration), 10 μM ZnCl2, and 10% glycerol. After washing steps, 20 uL sample buffer (1X laemmli buffer diluted from 4X in 8 M urea) was added to each sample. Samples were incubated for 30 minutes at 37C and resolved by denaturing polyacrylamide gel electrophoresis. Proteins were transferred to either nitrocellulose or PVDF membranes by semi-dry transfer using the Bio-Rad Trans-Blot Turbo transfer system. Membranes were blocked in 5% milk for 30 minutes, and proteins were imaged by immuno-blotting using the Thermo Super Signal Pico PLUS chemiluminescent substrate.

The recombinant N-terminally GST-tagged KLF4-ZF domain was expressed in BL21

Rosetta (DE3) bacterial cells using 1 mM IPTG induction performed overnight at 16C. 200 mL cell culture was collected and incubated in 10 mL

lysis buffer (PBS, 0.1% NP-40, 1 mM DTT, CPIs, 10 μM ZnCl2, 2.5 μM MgCl2, 0.25 mg/ml lysozyme, 2 uL benzonase [>500 U]) on ice for ten minutes. Cells were lysed by 30 second rounds of sonication followed by thirty seconds incubation on ice until the lysate cleared. Lysates were centrifuged at 4600 g for thirty minutes at 4C and the soluble supernatant was immediately added to 150 uL glutathione-agarose beads that had been prewashed 1X with DBB (Pierce). Ethidium bromide was added to the supernatant to a final concentration of 20 μg/mL, and lysates were incubated with beads for 1 hour while rotating end-over-end at 4C. Beads were washed six times with DBB (1 M NaCl) and three times with PBS. GST-tagged proteins were eluted from the beads with 50 mM reduced glutathione in PBS. Eluted proteins were dialyzed into PBS (+10 μM ZnCl2) twice over two hours at 4C before a final dialysis overnight into PBS (+10 μM ZnCl2) at 4C using Slide-A-Lyzer dialysis cassettes (10,000 MWCO, 0.1-0.5mL volume, Thermo). NP-40 was added to 0.1% and DDT was added to 1mM, while NaCl was adjusted to 400 mM in PBS for storage. Protein concentration was measured by UV absorbance at 280 nM and by Bradford assay. KLF4-ZF experiments were repeated after dialyzing GST-KLF4-ZF into PBS with no additions, and similar results were observed.

The PHF10 C-terminal double PHD finger domain and the SMARCB1 N-terminal winged-helix domain (WHD) were N-terminally GST-tagged during ligation-based cloning, expressed in BL21 Rosetta (DE3) bacteria, and used for H3 peptide pulldowns as described previously[39]. Antibodies and dilutions used in this study are listed in Supplementary Table 2, and raw gel images are shown in Supplementary Fig. 11.

Electrophoretic mobility shift assays (EMSA)

For EMSA experiments probing for endogenous protein from HeLa lysates, only PBB was used as the incubation buffer. DNA oligos were diluted in PBB as described above and incubated with ~0.75 μg/uL HeLa nuclear lysate for 15 minutes at room temperature. Binding reactions were then resolved by native polyacrylamide gel electrophoresis in 0.5X TBE running buffer. Protein transfer and immunoblotting was performed as described above. For EMSAs using purified recombinant SP3, the oligo dilution and protein binding steps were performed in protein binding buffer supplemented with 10 ng/uL BSA (final concentration), 10 μM ZnCl2, and 10% glycerol. 100 ng SP3 protein per

reaction was used for the recombinant SP3 EMSA. Recombinant proteins used in this study are listed in Supplementary Table 2.

For KLF4-ZF EMSA experiments, proteins were prepared with DNA based on molar concentration as shown in Supplementary Fig. 3C in PBS. Binding reactions were incubated for 15 minutes at room temperature and resolved on a 1% agarose gel using 0.5X TAE running buffer. The agarose gel was washed for 5 minutes in distilled water, stained for thirty minutes with 0.5 μg/mL ethidium bromide, and destained for 5 minutes in distilled water before imaging.

Fluorescence polarization and fluorescence intensity assays

Fluorescence polarization and intensity experiments were performed largely as described previously[40,41]. SP/KLF wild-type oligonucleotides were ordered with a 5' Cy5 fluorescent label on each strand (IDT). Labelled oligonucleotides were annealed as described above and diluted to 2 nM in PBS. Recombinant KLF4-ZF was diluted using PBS to a three-fold dilution series of ten concentrations from 0.3 nM to 6 μM in a Greiner black, 96-well non-binding microplate. Recombinant KLF4-ZF and labelled oligonucleotides were mixed at a ratio of 1:1 in a final volume of 200 uL, with labelled oligonucleotides at a final concentration of 1 nM and recombinant KLF4-ZF at a final concentration range of 0.15 nM to 3 μM. Binding reactions were incubated for 20 minutes at room temperature, and fluorescence polarization and intensity were measured at 25C on a Tecan Spark 10M microplate reader. Baseline polarization was calibrated based on protein-free reference samples to 50 mP. Wells were measured with 200 flashes per well and a 1 second settling time per sample. Binding assays were performed in triplicate, and binding parameters were calculated and plotted as described above using a Hill-like function.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository[42] with the dataset identifier PXD007132. All other relevant data are available from the authors.

Results

Benchmarking K_d^{app} measurements with the SP/KLF motif

We first affinity purified nuclear proteins from isolated nuclear lysates using a series of ten pulldowns with different concentrations of oligonucleotide baits coupled to streptavidin-sepharose beads. Proteins binding at each titration point were digested to tryptic peptides and isobarically labelled with the 10-plex TMT system[18]. Labelled peptides were combined and measured in a single SPS-MS3 mass spectrometry run[24,25] (Fig. 1A). Critically, the highest titration point (labelled with TMT131) represented a pulldown at micromolar concentration. We assumed that proteins with nanomolar range apparent dissociation constants (K_d^{app}~1-500 nM) would exhibit saturated binding at this concentration. Thus, for each DNA concentration we calculated the bound fraction for each individual protein compared to the TMT131 reporter ion signal (Fig 1B, Supplementary Fig. 1). K_d^{app} values were determined independently for each protein by fitting the parameters of a Hill-like curve using the known DNA concentrations and the observed fraction bound (Fig. 1A). We filtered out background or non-specific proteins based on the quality of fit of the Hill curve, under the assumption that background proteins would show randomly distributed ratios near 1:1 for all titration points. As such, only proteins fitting a Hill-like curve with an r-squared value greater than 0.95 were kept for downstream analysis. As each bait profiled required a set of ten pulldowns conducted in triplicate or duplicate for fitting, we utilized a filter plate system to increase throughput[15]. We note that this system is amenable to automation in future high-throughput studies.

As a benchmark case, we quantitatively profiled the nuclear protein binding landscape of the well-characterized SP/KLF consensus GC-box[15]. Oligo depletion was essentially complete after bead immobilization (Supplementary Fig. 2A). Binding of canonical SP/KLF factors was strongly specific for the designed SP/KLF motif oligonucleotide, with essentially no background binding observed for a mutated SP/KLF motif (Supplementary Fig. 2B). Using an r-squared value of 0.95 as a filtering criterion, we observed low coefficients of variation for fitted K_d^{app} values (Fig. 1C). We estimated the K_d^{app} value for canonical SP/KLF binding factor SP1 as ~38 nM and confirmed this result in lysates with gel-based assays (Fig. 1D, Supplementary Fig. 2C). Furthermore, we estimated K_d^{app} values for a number of other SP/KLF family factors including

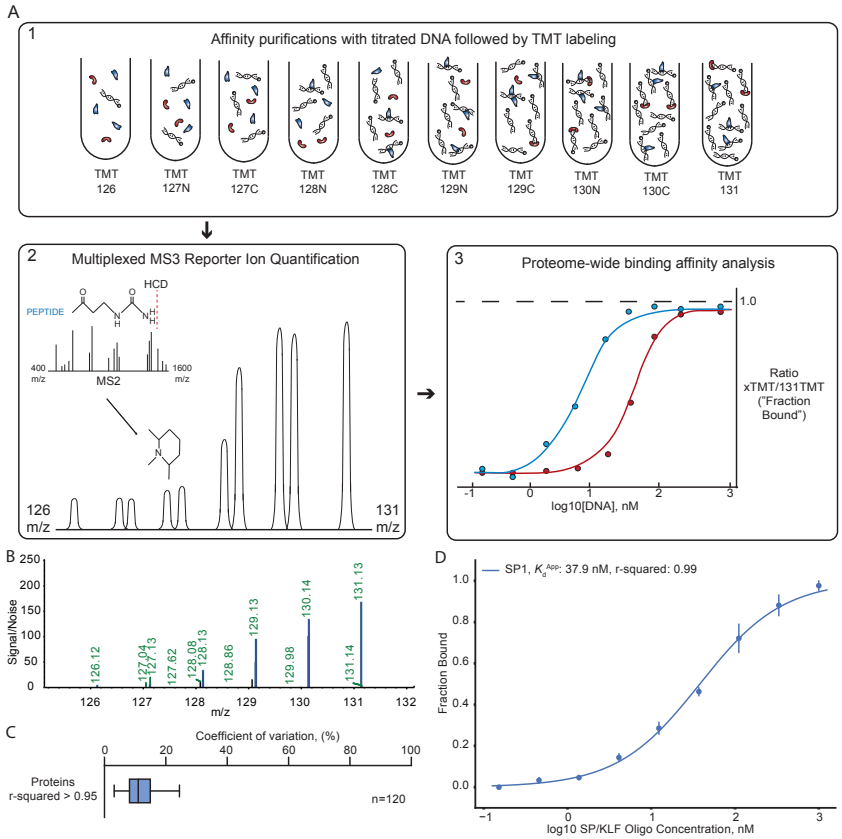


Figure 1. Benchmarking protein-DNA K_d^{App} measurements with the SP/KLF consensus motif

A A titration series of a known concentration of bait is used for affinity purification of proteins from nuclear lysates. Bound proteins are digested with trypsin, isobarically labelled with TMT reagent, and analyzed by mass spectrometry. Quantification of binding interactions yields a Hill-like curve, as described in the Methods, which can be used to calculate the K_d^{App}.

B SPS-MS3 TMT reporter ion spectrum of an example SP1 peptide. Only the low m/z range of the MS3 spectrum, where the TMT reporter ions are observed, is displayed for clarity. Plotted on the y-axis are signal-to-noise values measured in the orbitrap at 60,000 resolution.

C Boxplot analysis of all coefficients of variation for fitted K_d^{App} values identified using the SP/KLF consensus motif with r-squared values > 0.95. The box represents the quartiles of the data, while the whiskers represent the range of 1.5 IQRs. The center line is the median of the distribution.

D Hill-like curve identified for SP1 binding to the consensus SP/KLF GC-box motif. Binding curves were generated by fitting the parameters of the Hill equation including K_d^{App}. Each data point is the mean of three experiments (n=3), and the error bars represent the standard error of the mean.

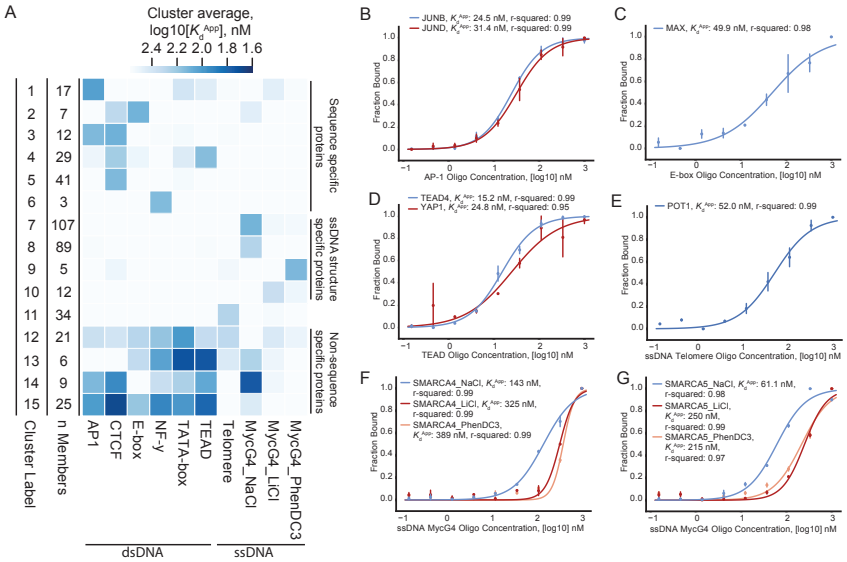


Figure 2. A motif survey identifies SWI/SNF and ISWI factors binding to G4-quadruplex structures

A Heatmap analysis K_d^{App} binding profiles for all dsDNA and ssDNA sequences and experiments. Proteins were clustered (15 clusters) using hierarchical agglomerative clustering. The heatmap is colored by the average log10(K_d^{App}) value of the cluster per bait. Cluster labels and number of proteins per cluster (n) are listed in columns to the left of the heatmap.

B-G K_d^{App} binding curves for canonical and unreported binding proteins for some example dsDNA and ssDNA motifs.

B K_d^{App} binding curve for AP-1 dsDNA motif and dimeric binding factors JUNB (Cluster 15) and JUND (Cluster 3).

C K_d^{App} binding curve for E-box dsDNA motif and dimeric binding factor MAX (Cluster 2).

D K_d^{App} binding curve for TEAD dsDNA motif and dimeric binding factors TEAD4 (Cluster 4) and YAP1 (Cluster 4).

E K_d^{App} binding curve for the telomere ssDNA motif (four repeats) and binding factor POT1 (Cluster 11).

F K_d^{App} binding curve for the mycG4 ssDNA motif in NaCl (G4-permissive), LiCl (G4-nonpermissive), and PhenDC3 (G4-ligand) binding conditions and SWI/SNF binding factor SMARCA4 (Cluster 7).

G K_d^{App} binding curve for the mycG4 ssDNA motif in NaCl, LiCl, and PhenDC3 binding conditions and ISWI binding factor SMARCA5 (Cluster 7).

For B-G), binding curves were generated by fitting the parameters of the Hill equation including K_d^{App}. Each data point is the mean of two experiments (n=2), and the error bars represent the standard error of the mean.

SP3 and KLF4. Interestingly, purified recombinant SP3 exhibited a substantially lower K_d in gel-shift assays indicating a higher affinity compared to mass

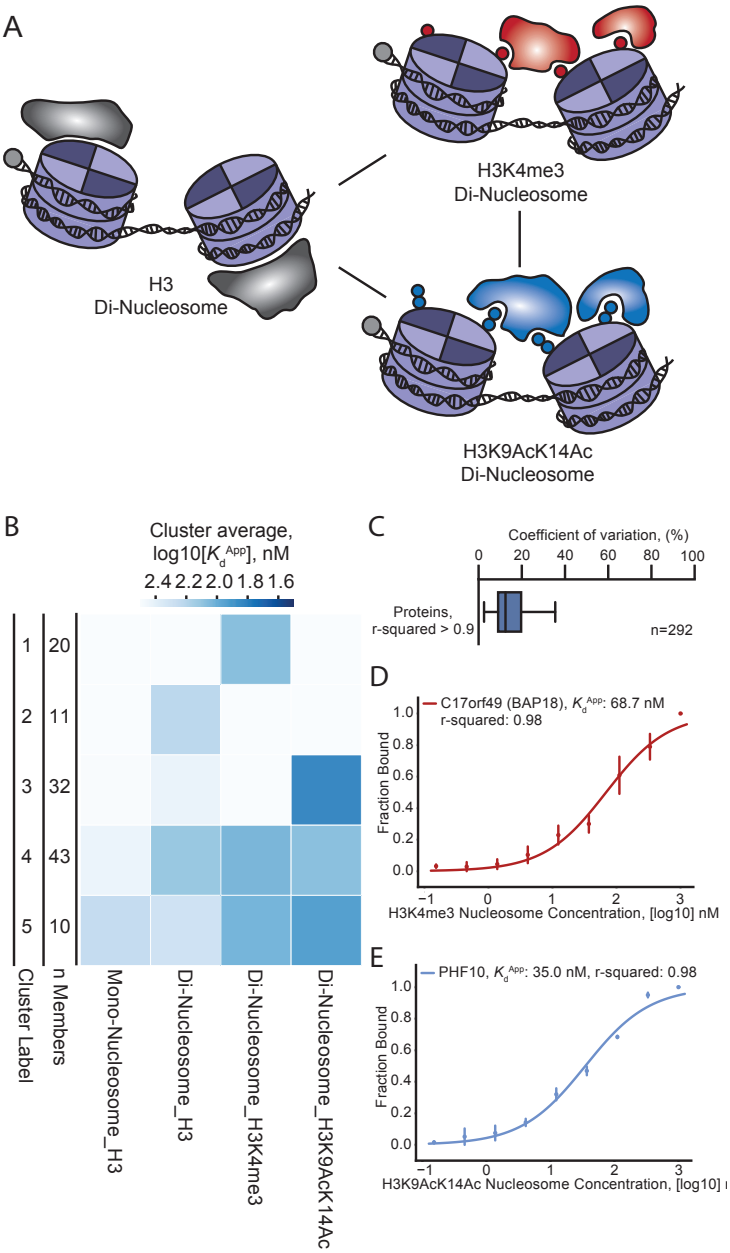


Figure 3. Quantitative analysis of modified di-nucleosome interactions

A Schematic representation of K_d^{App} di-nucleosome and modified di-nucleosome study design.
B Heatmap analysis of K_d^{App} binding profiles for all nucleosome, di-nucleosome, and modified di-nucleosome experiments. Proteins were clustered (5 clusters) using hierarchical agglomerative clustering. The heatmap is colored by the average log10(K_d^{App}) value of the cluster per bait. Cluster labels and number of proteins per cluster (n) are listed in columns to the left of the heatmap.
C Boxplot analysis of all coefficients of variation for fitted K_d^{App} values identified in nucleosome experiments with r-squared values > 0.90. The box represents the quartiles of the data, while the whiskers represent the range of 1.5 IQRs. The center line is the median of the distribution.
D-E Example K_d^{App} binding curves for binding proteins of H3K4me3 and H3K9AcK14Ac di-nucleosomes.
D K_d^{App} binding curve for H3K4me3 modified di-nucleosomes and binding factor C17orf49 (Cluster 1).
E K_d^{App} binding curve for H3K9AcK14Ac modified di-nucleosomes and binding factor PHF10 (Cluster 3).

Binding curves were generated by fitting the parameters of the Hill equation including K_d^{App}. Each data point is the mean of three experiments (n=3), and the error bars represent the standard error of the mean.

spectrometry-based measurements from nuclear lysates (Supplementary Fig. 2D). However, this shift was abrogated when recombinant SP3 was spiked into nuclear lysates, where recombinant and endogenous SP3 showed comparable binding curves. Similarly, we observed a lower K_d value for the SP/KLF motif and purified recombinant KLF4-ZF domain using fluorescence polarization and fluorescence de-quenching assays compared to those measured for endogenous KLF4 in lysates by mass spectrometry and gel-based assays (Supplementary Fig. 3)[40,41]. This clearly suggests competitive inhibition between proteins in the nuclear environment as has been reported for SP1-KLF4 and SP1-KLF16 among other SP/KLF factors[43]. As such, this complex interplay between DNA binding proteins indicates K_d^{app} measurements will be useful information for uncovering interactions within transcriptional networks as they exist in the *in vivo* nuclear environment.

To further characterize the specificity and sensitivity of this assay, we conducted a series of competition experiments with free (unbiotinylated) wild-type and mutated SP/KLF oligonucleotides. In agreement with western blot analysis (Supplementary Fig. 2B), known sequence-specific SP/KLF transcription factors showed no measurable binding to the mutated SP/KLF oligonucleotide (Supplementary Fig. 4A). In competition experiments, these sequence-specific factors were similarly not competed away from immobilized

wild-type oligonucleotides by free mutated oligonucleotides (Supplementary Fig. 4A). Yet, importantly, K_d^{app} values for sequence-specific proteins estimated by applying the Cheng-Prusoff correction to IC_{50} values from competition experiments (Supplementary Fig. 4B) were highly correlated to K_d^{app} values estimated with immobilized oligonucleotides at a near 1:1 ratio (Supplementary Fig. 4C). In contrast, non-sequence specific proteins bound to immobilized wild-type or mutated SP/KLF oligos with high correlation (Supplementary Fig. 4A). Non-sequence specific proteins were also competed from immobilized wild-type oligonucleotides by either wild-type of mutated free oligonucleotide. However, non-sequence specific proteins displayed lower K_d^{app} values on average to either free oligonucleotide compared to immobilized oligonucleotides in competition experiments (Supplementary Fig. 4C). We attribute this to free oligonucleotides having two available blunt ends acting as substrates for some non-sequence specific proteins, while immobilized oligonucleotides only have one sterically free blunt end. This is an important consideration for future studies comparing free v. immobilized oligonucleotides and sequence-specific v. non-specific DNA binding factors.

The K_d^{app} values we measured did not significantly correlate with absolute protein abundance, demonstrating that the K_d^{app} values we measured were not biased by protein abundance (Supplementary Fig. 4D, Supplementary Data 2)[28,44]. An additional analysis taking possible phospho-post-translational modifications into account also showed practically no difference with our original result, suggesting such phospho-modifications, even if stoichiometric *in vivo*, may not be easily identified without specific enrichment methods (Supplementary Fig. 4E)[45]. Finally, we calculated JASPAR transcription factor binding profile motif scores for sequence-specific proteins with JASPAR motifs and correlated them with the K_d^{app} values we measured (Supplementary Fig. 5). We observed significant correlation, demonstrating that our assay produced results consistent with motif-based binding landscape models. Summarily, on these bases we concluded that our assay generated reliable K_d^{app} measurements for protein-DNA interactions in this SP/KLF benchmark case.

Chromatin remodelers quantitatively prefer G-quadruplexes

Next, we conducted a larger assay of eight different dsDNA or ssDNA (Fig. 2A, Supplementary Table 1) sequences representing canonical, well-characterized biological motifs (AP-1, CTCF, E-box, NF-Y, TATA, TEAD, the human ssDNA telomere repeat, and the c-Myc promoter Pu27 G4-quadruplex forming ssDNA sequence [mycG4][46]). We observed numerous homo- and hetero-multimeric transcription factors and DNA-binding proteins binding to their canonical motif in readily distinguishable clusters, including JUNB/JUND, MAX, POT1, and TEAD/YAP (Fig. 2A-E). More specifically, we were surprised and intrigued to see subunits of many chromatin modifying complexes binding to the mycG4 sequence in G4-permissive NaCl-based binding buffer, including SWItch/Sucrose Non-Fermentable (SWI/SNF) and Imitation SWI (ISWI) subunits (Fig. 2F-G). This immediately suggested the possibility that some chromatin remodeling and modifying enzymes specifically recognize DNA G4 structures. To further characterize these interactions, we performed an additional set of assays using either 150 mM LiCl buffer, which destabilizes G-quadruplex structures, or 150 mM NaCl binding buffer supplemented with 20 μM PhenDC3, which stabilizes G4-structures (Fig. 2A). Both LiCl and G-quadruplex stabilizing small molecules were shown previously to inhibit G4-RNA binding of Polycomb repressive complex 2 (PRC2) [47], which binds RNA G-quadruplexes through interfaces on EZH2 and EED subunits and an RNA recognition motif in the SUZ12 subunit[48,49]. Here, we observed that K_d^{app} values measured for the mycG4 sequence in LiCl and PhenDC3 binding conditions were indeed positively correlated (Supplementary Fig. 6A). LiCl or PhenDC3 treatment quantitatively and significantly increased K_d^{app} values for both SWI/SNF and ISWI subunits and for PRC2 and Nucleosome remodeling and deacetylase (NuRD) subunits (Fig. 2F-G, Supplementary Fig. 6B-C). Among these, PRC2 and NuRD subunits bound with ~100 nM K_d^{app}s to the mycG4 sequence, similar to the recently reported binding affinity of PRC2 to RNA G-quadruplexes[47,48]. Further supporting this finding, we detected enrichment of induced G-quadruplex structures in both SWI/SNF, PRC2, and NuRD subunit binding sites in a variety of ENCODE ChIP-seq datasets, consistent with our *in vitro* binding data (Supplementary Fig. 6D)[33-35]. We also verified this finding by western blot (Supplementary Fig. 7A-C). Finally, we were interested in identifying potential direct G4 DNA-binding subunits from these

chromatin remodeling and modifying complexes. We adapted a formaldehyde cross-linking approach to identify individual peptides immediately proximal to the mycG4 DNA bait (Supplementary Fig. 8A)[50]. Identified peptides from a known DNA-binding complex were proximal to the DNA-binding channel, indicating the reliability of the approach (Supplementary Fig. 8B, PDB: 1JEY [http://dx.doi.org/10.2210/pdb1JEY/pdb])[51]. We identified a peptide from the N-terminal winged helix-like domain (WHD) of SMARCB1 as the most significantly enriched for direct G4 binding of SWI/SNF, ISWI, PRC2, and NuRD subunits (Supplementary Fig. 8C-E). Affinity purification experiments with the recombinant SMARCB1-WHD verified that the SMARCB1-WHD specifically recognizes the mycG4 bait compared to a variety of control baits while the double PHD finger of PHF10, another SWI/SNF accessory subunit, does not (Supplementary Fig. 8F)[52]. Overall, these *in vitro* analyses provide biochemical support for the hypothesis that some chromatin remodelers and modifiers can bind G4 DNA sequences in a G-quadruplex preferential manner.

K_d^{app} estimation using nucleosome substrates

Finally, to show that this binding assay is compatible not only with nucleic acids but also with nucleoprotein complexes, we performed a set of experiments using mono-nucleosomes, di-nucleosomes, and modified di-nucleosomes (Fig. 3A, Supplementary Fig. 9A-D). We identified a number of protein-di-nucleosome interactions that were either specific to or modulated by either H3K4me3 or H3K9AcK14Ac (Fig 3B, Fig. 3D-E). Across all baits, coefficients of variation were reasonably low indicating good data quality (Fig 3C). We observed SWI/SNF and ISWI binding with significantly lower K_d^{app} to H3K9AcK14Ac-containing di-nucleosomes compared to unmodified di-nucleosomes (Supplementary Fig. 10A/D). Indeed, many accessory SWI/SNF subunits were uniquely identified with the H3K9AcK14Ac di-nucleosome bait. We identified one SWI/SNF subunit, PHF10, which harbors a C-terminal double PHD finger (DPF) and may preferentially interact with H3K9AcK14Ac compared to unmodified H3 or H3K4me3 as has been reported for another DPF domain[53]. PHF10 shows one of the lowest K_d^{app} values of all identified SWI/SNF subunits (~35 nM, Fig 3E) for H3K9AcK14Ac di-nucleosomes and was similarly identified exclusively with the H3K9AcK14Ac di-nucleosome. To further investigate the H3K9AcK14Ac specificity of PHF10, we used bacterial

lysates expressing the recombinant GST-tagged DPF domain of PHF10 in histone H3 peptide pull-down experiments (Supplementary Fig. 10B). This experiment clearly confirmed that recombinant PHF10-DPF binding to H3 is strongly agonised by H3K9AcK14Ac in contrast to the SMARCB1-WHD, as has been reported for the bromodomain of Swi2/Snf2 and the DPF domains of SWI/SNF accessory subunits DPF2 and DPF3 (Supplementary Fig. 10C)[54-56]. Thus, PHF10 represents an additional high affinity H3 acetylation reader in the human SWI/SNF complex. Intriguingly, PHF10-DPF seems to be repelled by H3K4me3, and our data suggests this might confer a lower affinity "fine-tuning" on PHF10-SWI/SNF nucleosome binding. However, even in the absence of the acetylation specificity conferred by SWI/SNF acetylation reader domains, we observed that catalytic SWI/SNF subunits SMARCA2 and SMARCA4 still engage in relatively high affinity interactions with H3K4me3 modified and unmodified di-nucleosomes (Supplementary Fig. 10A/D). Similarly, we observed relatively high affinity interactions between modified or unmodified di-nucleosomes and ISWI catalytic subunit SMARC5 and ISWI accessory subunits including BPTF, C17orf49 (Fig. 3E), BAZ1A, and BAZ1B. The binding patterns we observe between different di-nucleosome baits are complex and highlight an important point and a unique benefit of measuring apparent affinities from complex lysates: we measure the average binding profile over what is likely a pool of heterogeneous multimeric protein complexes. For example, we observe catalytic subunit SMARCA5 binding with highest affinity to H3K9AcK14Ac di-nucleosomes compared to unmodified or H3K4me3 di-nucleosomes (Supplementary Fig. 10D). Yet, accessory subunits BPTF, C17orf49, BAZ1A, and BAZ1B exhibit binding curves that are both similar to and distinct from the binding curve of SMARCA5 depending on the accessory subunit and di-nucleosome pair in question, suggesting these subunits regulate differential H3 modification-specific binding (Supplementary Fig. 10D). Moreover, these ISWI accessory subunits, along with SMARCA5, form at least three unique and independent protein complexes (the NuRF, ACF, and WICH complexes)[57]. More generally, deconvoluting the individual contributions of different subcomplexes with recombinant systems is indisputably critical (for example, as with SMARCB1 and its G-quadruplex interaction and PHF10-H3K9AcK14Ac binding in SWI/SNF). In the future, a data-driven strategy for assessing the contributions of shared subunits within

independent subcomplexes might involve a combinatorial approach utilizing many DNA or modified nucleosome baits as was demonstrated, for example, with the stoichiometry of different SET1/MLL subcomplexes[28]. Additionally, we argue that the ability to measure binding affinities for multimeric complexes in the complex environment of the nucleus offers an important holistic, system-wide view.

Discussion

By measuring protein-DNA and protein-nucleosome K_d^{app} values via mass spectrometry, we provide another avenue for extending protein-nucleic acid interaction proteomics beyond comparative, semi-quantitative workflows. From a systems biology perspective, absolute binding affinities within the nuclear environment create a quantitative link between transcription factor expression and target gene regulation via the TF-DNA interaction. However, our data implies a situation where DNA sequences, structures, and chromatinized DNA integrate signals from a variety of binding partners at a spectrum of biologically relevant affinities. Indeed, we observe that both DNA structures, including G-quadruplexes, and (modified) nucleosomes engage in high affinity binding interactions with various independent chromatin remodeling and modifying complexes. Of note, we observe SWI/SNF as a complex that recognizes both DNA structure and histone modification state, and we identify SMARCB1 and PHF10, respectively, as specific "readers" that contribute to these high affinity substrate recognitions. While targeted biochemical analysis reveals direct mediators of specific substrate interactions within a complex, combining these analyses with affinity measurements in the context of multimeric assemblies in complex lysates takes into consideration the possibility of larger molecular regulatory networks. This work demonstrates an assay for revealing not only these regulatory interactions but also their absolute affinities and thereby profiling the proteome-wide quantitative binding landscape of nuclear proteins for DNA oligonucleotides and nucleosome complexes.

References

1 Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712-723, doi:10.1016/j.cell.2015.09.053 (2015).

2 Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509, doi:10.1038/nature22366 (2017).

3 Drew, K. *et al.* Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol Syst Biol* **13**, 932, doi:10.15252/msb.20167490 (2017).

4 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355, doi:10.1038/nature19949 (2016).

5 Smits, A. H. & Vermeulen, M. Characterizing Protein-Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends Biotechnol* **34**, 825-834, doi:10.1016/j.tibtech.2016.02.014 (2016).

6 Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-1664, doi:10.1126/science.1069492 (2002).

7 Bantscheff, M. *et al.* Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nat Biotechnol* **29**, 255-265, doi:10.1038/nbt.1759 (2011).

8 Sharma, K. *et al.* Proteomics strategy for quantitative protein interaction profiling in cell extracts. *Nat Methods* **6**, 741-744, doi:10.1038/nmeth.1373 (2009).

9 Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346**, 1255784, doi:10.1126/science.1255784 (2014).

10 Mateus, A., Maatta, T. A. & Savitski, M. M. Thermal proteome profiling: unbiased assessment of protein state through heat-induced stability changes. *Proteome Sci* **15**, 13, doi:10.1186/s12953-017-0122-4 (2016).

11 Stormo, G. D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **11**, 751-760, doi:10.1038/nrg2845 (2010).

12 Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233-237, doi:10.1126/science.1131007 (2007).

13 Nordhoff, E. *et al.* Rapid identification of DNA-binding proteins by mass spectrometry. *Nat Biotechnol* **17**, 884-888, doi:10.1038/12873 (1999).

14 Spruijt, C. G., Baymaz, H. I. & Vermeulen, M. Identifying specific protein-DNA interactions using SILAC-based quantitative proteomics. *Methods Mol Biol* **977**, 137-157, doi:10.1007/978-1-62703-284-1_11 (2013).

15 Hubner, N. C., Nguyen, L. N., Hornig, N. C. & Stunnenberg, H. G. A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. *J Proteome Res* **14**, 1315-1329, doi:10.1021/pr5009515 (2015).

16 Lau, H. T., Suh, H. W., Golkowski, M. & Ong, S. E. Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics. *J Proteome Res* **13**, 4164-4174, doi:10.1021/pr500630a (2014).

17 Makowski, M. M. *et al.* An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations. *Proteomics* **16**, 417-426, doi:10.1002/pmic.201500327 (2016).

18 McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem* **84**, 7469-7478, doi:10.1021/ac301572t (2012).

19 Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2**, 1896-1906, doi:10.1038/nprot.2007.261 (2007).

20 Dyer, P. N. *et al.* Reconstitution of nucleosome core particles from recombinant histones and DNA. *Methods Enzymol* **375**, 23-44 (2004).

21 Bartke, T. *et al.* Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143**, 470-484, doi:10.1016/j.cell.2010.10.012 (2010).

22 Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol* **276**, 19-42, doi:10.1006/jmbi.1997.1494 (1998).

23 Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *J Mol Biol* **327**, 85-96 (2003).

24 Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8**, 937-940, doi:10.1038/nmeth.1714 (2011).

25 McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **86**, 7150-7158, doi:10.1021/ac502040v (2014).

26 Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923-925, doi:10.1038/nmeth1113 (2007).

27 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

28 van Nuland, R. *et al.* Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Mol Cell Biol* **33**, 2067-2077, doi:10.1128/MCB.01742-12 (2013).

29 Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146-1159, doi:10.1016/j.cell.2013.02.004 (2013).

30 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342, doi:10.1038/nature10098 (2011).

31 Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359-362, doi:10.1038/nmeth.1322 (2009).

32 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).

33 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

34 Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**, D726-732, doi:10.1093/nar/gkv1160 (2016).

35 Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**, 877-881, doi:10.1038/nbt.3295 (2015).

36 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102. Article published online before print in May 2002 (2002).

37 Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34, doi:10.1002/0471250953.bi1112s47 (2014).

38 Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423-3424, doi:10.1093/bioinformatics/btr539 (2011).

39 Vermeulen, M. *et al.* Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142**, 967-980, doi:10.1016/j.cell.2010.08.020 (2010).

40 Anderson, B. J., Larkin, C., Guja, K. & Schildbach, J. F. Using fluorophore-labeled oligonucleotides to measure affinities of protein-DNA interactions. *Methods Enzymol* **450**, 253-272, doi:10.1016/S0076-6879(08)03412-5 (2008).

41 Hieb, A. R., D'Arcy, S., Kramer, M. A., White, A. E. & Luger, K. Fluorescence strategies for high-throughput quantification of protein interactions. *Nucleic Acids Res* **40**, e33, doi:10.1093/nar/gkr1045 (2012).

42 Vizcaino, J. A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* **41**, D1063-1069, doi:10.1093/nar/gks1262 (2013).

43 Lomberk, G. & Urrutia, R. The family feud: turning off Sp1 by Sp1-like KLF proteins. *Biochem J* **392**, 1-11, doi:10.1042/BJ20051234 (2005).

44 Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**, D142-147, doi:10.1093/nar/gkt997 (2014).

45 Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9**, 4632-4641, doi:10.1002/pmic.200900398 (2009).

46 You, H., Wu, J., Shao, F. & Yan, J. Stability and kinetics of c-MYC promoter G-quadruplexes studied by single-molecule manipulation. *J Am Chem Soc* **137**, 2424-2427, doi:10.1021/ja511680u (2015).

47 Wang, X. *et al.* Targeting of Polycomb Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines. *Mol Cell* **65**, 1056-1067 e1055, doi:10.1016/j.molcel.2017.02.003 (2017).

48 Long, Y. *et al.* Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *Elife* **6**, doi:10.7554/eLife.31558 (2017).

49 Kasinath, V. *et al.* Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science* **359**, 940-944, doi:10.1126/science.aar5700 (2018).

50 Ponicsan, S. L. *et al.* The non-coding B2 RNA binds to the DNA cleft and active-site region of RNA polymerase II. *J Mol Biol* **425**, 3625-3638, doi:10.1016/j.jmb.2013.01.035 (2013).

51 Walker, J. R., Corpina, R. A. & Goldberg, J. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature* **412**, 607-614, doi:10.1038/35088000 (2001).

52 Choi, J. *et al.* A common intronic variant of PARP1 confers melanoma risk and mediates melanocyte growth via regulation of MITF. *Nat Genet* **49**, 1326-1335, doi:10.1038/ng.3927 (2017).

53 Dreveny, I. *et al.* The double PHD finger domain of MOZ/MYST3 induces alpha-helical structure of the histone H3 tail to facilitate acetylation and methylation sampling and modification. *Nucleic Acids Res* **42**, 822-835, doi:10.1093/nar/gkt931 (2014).

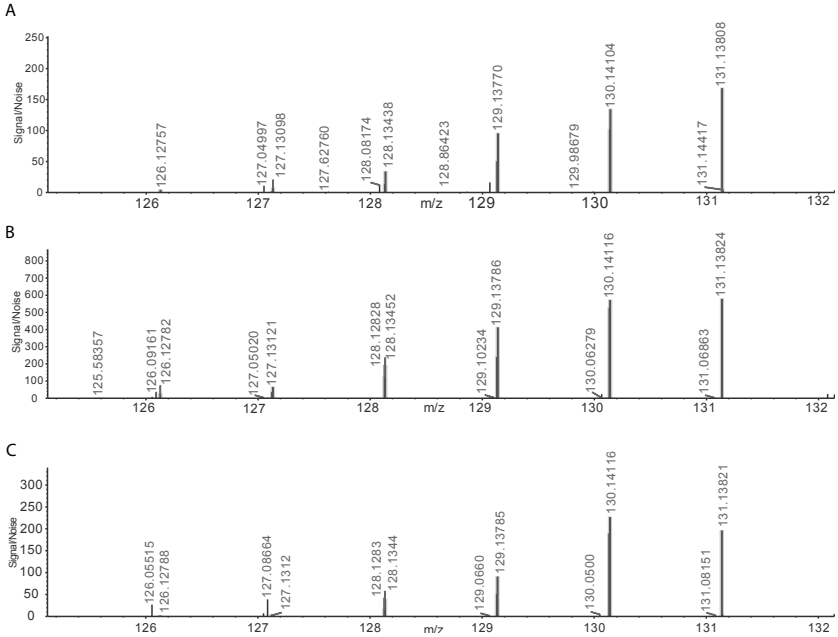
54 Awad, S. & Hassan, A. H. The Swi2/Snf2 bromodomain is important for the full binding and remodeling activity of the SWI/SNF complex on H3- and H4-acetylated nucleosomes. *Ann N Y Acad Sci* **1138**, 366-375, doi:10.1196/annals.1414.038 (2008).

55 Lange, M. *et al.* Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. *Genes Dev* **22**, 2370-2384, doi:10.1101/gad.471408 (2008).

56 Matsuyama, R. *et al.* Double PHD fingers protein DPF2 recognizes acetylated histones and suppresses the function of estrogen-related receptor alpha through histone deacetylase 1. *J Biol Chem* **285**, 18166-18176, doi:10.1074/jbc.M109.077024 (2010).

57 Erdel, F. & Rippe, K. Chromatin remodelling in mammalian cells by ISWI-type complexes--where, when and why? *FEBS J* **278**, 3608-3618, doi:10.1111/j.1742-4658.2011.08282.x (2011).

SUPPLEMENTARY FIGURES

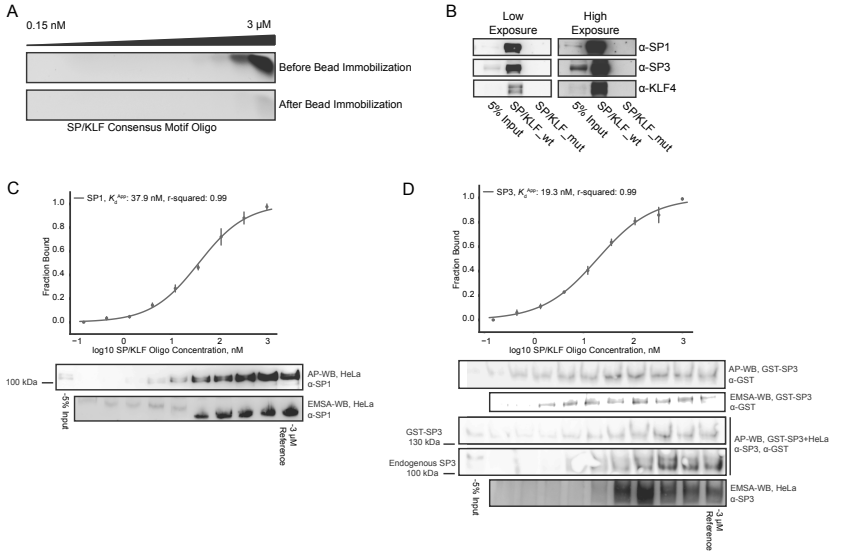


Supplementary Figure 1. Example MS3 TMT reporter ion spectra used for protein quantification

A Example MS3 TMT reporter ion spectrum of an identified SP1 peptide. Only the low m/z range of the MS3 spectrum, where the TMT reporter ions are observed, is displayed for clarity. Plotted on the y-axis are signal-to-noise values from the orbitrap at 60,000 resolution.

B Example MS3 TMT reporter ion spectrum of an identified SP3 peptide.

C Example MS3 TMT reporter ion spectrum of an identified KLF4 peptide.



Supplementary Figure 2. Gel-based validation of oligo depletion and protein binding to SP/KLF motif

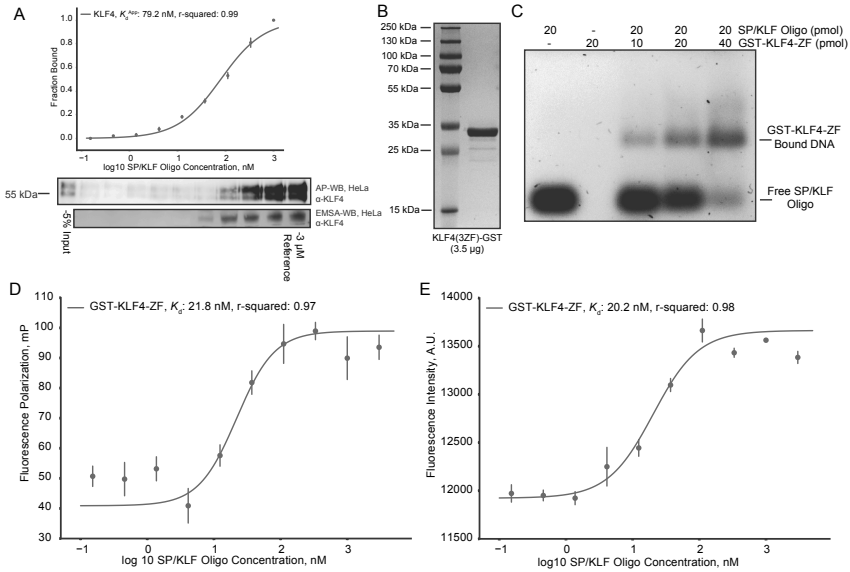
A Agarose gel electrophoresis of SP/KLF consensus oligo titration, both before and after binding to streptavidin-sepharose beads, to indicate depletion of the oligo by bead immobilization.

B Western blot analysis of canonical SP/KLF binding factors (SP1, SP3, and KLF4) binding to the SP/KLF wild-type oligonucleotide and not the SP/KLF mutated oligonucleotide.

C Validation of mass spectrometry binding profile for SP1 to the SP/KLF consensus oligo. The mass spectrometry binding curve is shown above. Affinity purification western blot and electrophoretic mobility shift assay (EMSA) data are shown below. Mass spectrometry data points and protein bands are approximately aligned on the vertical axis.

D Validation of mass spectrometry binding profile for SP3 to the SP/KLF consensus oligo. The mass spectrometry binding curve is shown above. Affinity purification western blot and electrophoretic mobility shift assay (EMSA) data of recombinant SP3 are shown below. Affinity purification western blot of recombinant SP3 spiked into HeLa lysate for pulldown analysis is shown next, with recombinant SP3 and endogenous SP3 separately noted. Finally, EMSA analysis of endogenous SP3 in HeLa nuclear lysate is shown at the bottom. Mass spectrometry data points and protein bands are approximately aligned on the vertical axis.

Binding curves were generated by fitting the parameters of the Hill equation including Kd^{App}. Each data point is the mean of three experiments (n=3), and the error bars represent the standard error of the mean.



Supplementary Figure 3. KLF4-ZF fluorescence polarization and fluorescence intensity assays

A Validation of mass spectrometry binding profile for KLF4 to the SP/KLF consensus oligo. The mass spectrometry binding curve is shown above. Affinity purification western blot and electrophoretic mobility shift assay (EMSA) data of endogenous KLF4 are shown below. Mass spectrometry data points and protein bands are approximately aligned on the vertical axis.

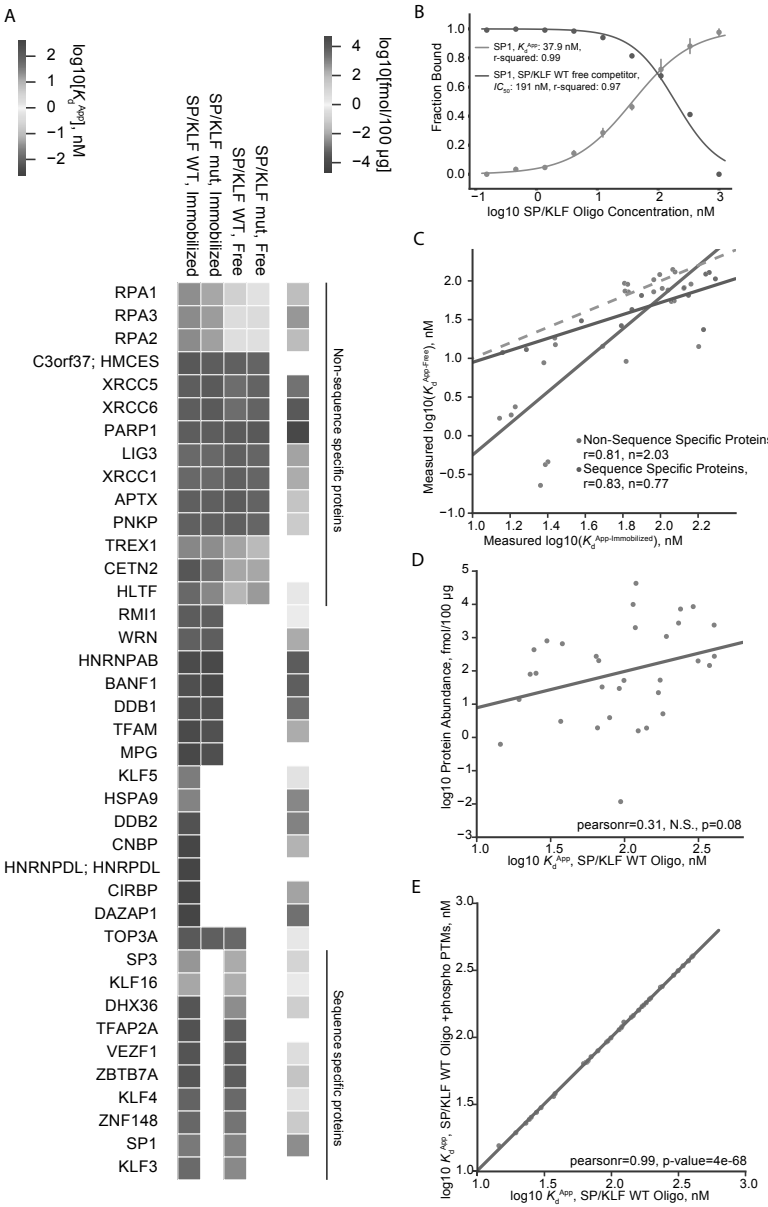
B Imperial protein stain following SDS-PAGE of 3.5 µg of purified GST-KLF4-ZF recombinant protein.

C Agarose gel electrophoretic mobility shift assay of GST-KLF4-ZF binding to the consensus SP/KLF oligonucleotide. Concentrations of either SP/KLF oligonucleotide or GST-KLF4-ZF are indicated above in picomoles.

D Fluorescence polarization assay of GST-KLF4-ZF binding to the Cy5 labelled SP/KLF oligonucleotide.

E Fluorescence intensity (de-quenching) assay of GST-KLF4-ZF binding to the Cy5 labelled SP/KLF oligonucleotide.

Binding curves were generated by fitting the parameters of the Hill equation including *K*dApp. Each data point is the mean of three experiments (n=3), and the error bars represent the standard error of the mean.



Supplementary Figure 4. *K*dApp calculated from competition experiment derived *IC*50 values correlates with *K*dApp measured for immobilized baits for sequence-specific proteins

A Heatmap analysis of log10-transformed *K*dApp values for SP/KLF consensus oligonucleotide experiments, including with mutated sequence and with competition experiments. *K*dApp values for competition experiments were calculated applying the Cheng-Prusoff correction to fit *IC*50 values as described in the Methods.

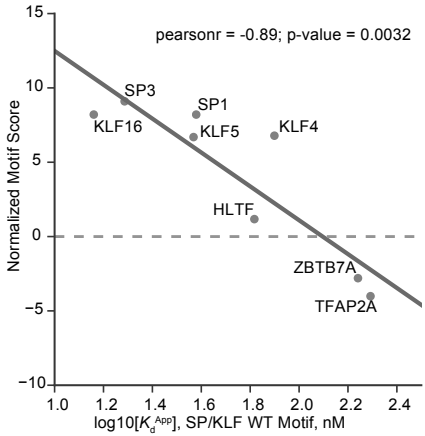
B Hill-like curve for SP1 binding to the consensus SP/KLF motif (*K*dApp) and for competition of SP1 by free SP/KLF motif(*IC*50).

C Regression analysis of *K*dApp values measured in experiments with immobilized baits compared to KdApp values calculated from *IC*50 values measured in competition experiments. Sequence-specific proteins (as indicated in panel A based on specific high affinity binding for the wild-type SP/KLF oligonucleotide), are plotted in red, as is the correlation between experiments. Non-sequence specific proteins (as indicated in panel A based on high affinity binding for both the wild-type SP/KLF oligonucleotide and the mutated SP/KLF oligonucleotide), are plotted in blue, as is the correlation between experiments. r is the pearson correlation, and n is the slope of the regression line.

D Regression analysis of *K*dApp values measured for the wild-type SP/KLF oligonucleotide in immobilized bait experiments compared to absolute protein abundance in nuclear lysates. Each data point represents a single protein. The correlation is not significant at a p-value of 0.05.

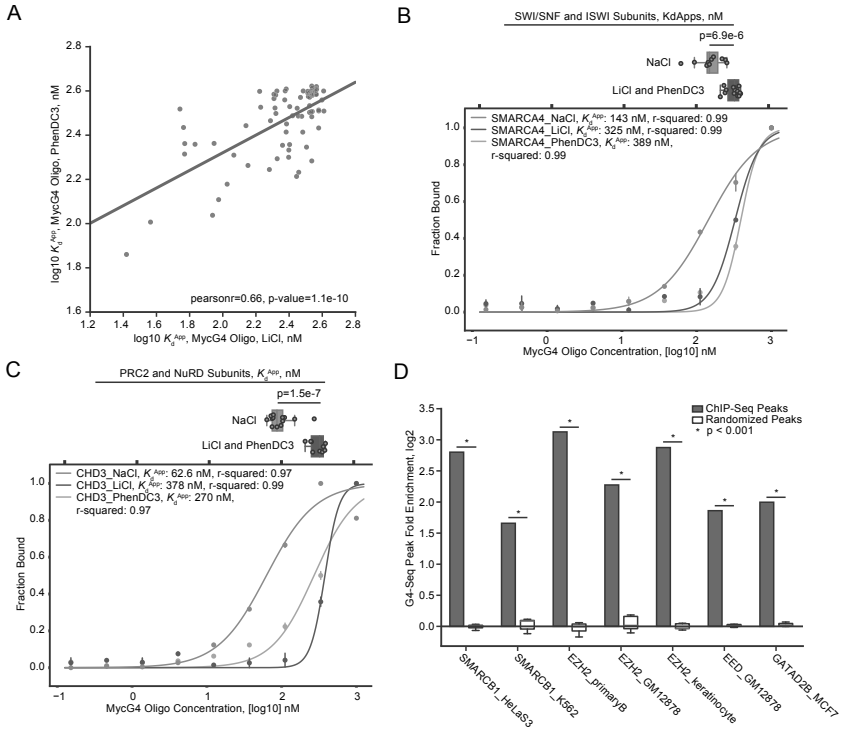
E Regression analysis of *K*dApp values measured for the wild-type SP/KLF oligonucleotide in immobilized bait experiments. Two analysis were performed and plotted, as described in the methods, either considering or not considering STY peptide phosphorylation. Each data point represents a single protein.

Binding curves were generated by fitting the parameters of the Hill equation including *K*dApp. Each data point is the mean of three experiments (n=3), and the error bars represent the standard error of the mean.



Supplementary Figure 5. Observed *K*dApp correlates with transcription factor motif score

Regression analysis between the log10-transformed *K*dApp value for sequence-specific SP/KLF wild-type oligonucleotide binding factors and the normalized JASPAR motif score for the SP/KLF consensus oligo as defined in the Methods. The grey dashed line, therefore, indicates the significance threshold for each individual factor.



Supplementary Figure 6. Chromatin modifying complexes bind with lower *K*dApp to the mycG4 sequence in G4-permissive conditions

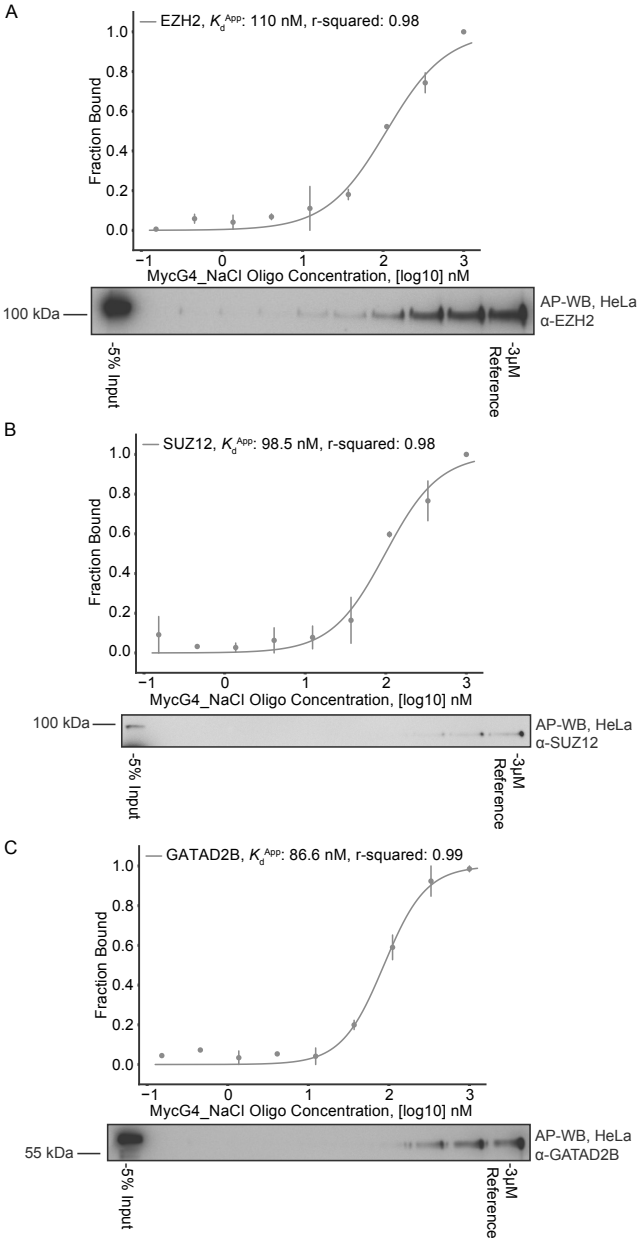
A Regression analysis of *K*d App values measured for the mycG4 ssDNA oligonucleotide in LiCl binding conditions compared to PhenDC3 binding conditions as described in the methods. Each data point represents a single protein.

B Binding curves for SMARCA4 (Fig. 2A, Cluster 7), a SWI/SNF catalytic subunit, are shown for the mycG4 oligonucleotide in NaCl (G4-permissive), LiCl (G4 non-permissive), and PhenDC3 (G4 ligand) binding conditions. Above are boxplots of all identified KdApp values SWI/SNF and ISWI subunits in these conditions, with LiCl and PhenDC3 grouped together given their similar overall effect (also shown in panel A).

C Binding curves for CHD3 (Cluster 7), a NuRD catalytic subunit, are shown for the mycG4 oligonucleotide in NaCl (G4-permissive), LiCl (G4 non-permissive), and PhenDC3 (G4 ligand) binding conditions. Boxplots and data points are colored as in panel B.

D Permutation based testing of G4-seq (generated in primary B cells) peak enrichment in ENCODE ChIP-seq peaks for SWI/SNF, PRC2, and NuRD subunits in various cell lines. ChIP-seq peaks were randomized 1000 times across the genome, and the distribution of enrichment values for randomized G4seq intersected peak counts, compared to the mean of peak randomizations, is displayed as a white boxplot with parameters as described previously. The enrichment of ChIP-seq peak enrichment compared to peak randomizations is indicated as a blue bar. An empirical p-value was calculated by comparing the distribution of randomized peak intersections with the true peak intersections.

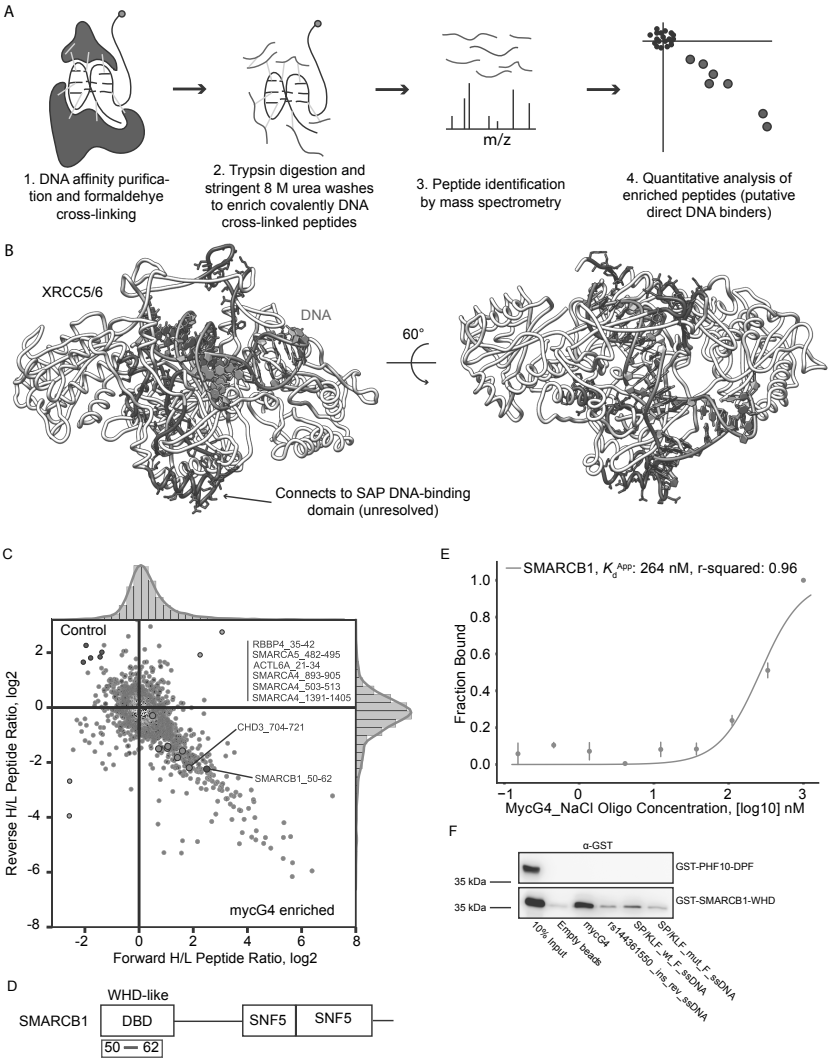
Boxplots are displayed such that the center line is the median of the distribution, the box represents the first and third quartile of the data, and the whiskers represent 1.5 inter-quartile ranges. Significance is indicated based on a two-sided t-test, with all measured values per sample, as indicated by the colored grouping, treated as sample populations.
Binding curves were generated by fitting the parameters of the Hill equation including *K*dApp. Each data point is the mean of two experiments (n=2), and the error bars represent the standard error of the mean.



Supplementary Figure 7. Western blot validation of PRC2 and NuRD binding to the mycG4 sequence

A Validation of mass spectrometry binding profiles for EZH2 (PRC2 subunit, Fig. 2A, Cluster 7) to the mycG4 consensus oligo. The mass spectrometry binding curve is shown above with affinity purification western blot data shown below. Mass spectrometry data points and protein bands are approximately aligned on the vertical axis.
B Validation of mass spectrometry binding profiles for SUZ12 (PRC2 subunit, Cluster 7) to the mycG4 consensus oligo. The mass spectrometry binding curve is shown above, with affinity purification western blot data shown below. Mass spectrometry data points and protein bands are approximately aligned on the vertical axis.
C Validation of mass spectrometry binding profiles for GATAD2B (NuRD subunit, Cluster 7) to the mycG4 consensus oligo. The mass spectrometry binding curves are shown above, with affinity purification western blot data shown below. Mass spectrometry data points and protein bands are approximately aligned on the vertical axis.

Binding curves were generated by fitting the parameters of the Hill equation including *K*dApp. Each data point is the mean of two experiments (n=2), and the error bars represent the standard error of the mean.



Supplementary Figure 8. The SMARCB1 winged helix domain directly binds to and prefers the mycG4 sequence

A Workflow for formaldehyde protein-DNA cross-linking experiments. Briefly, proteins were covalently cross-linked to DNA oligonucleotides. Proteins were digested and non-linked peptides were removed by stringent 8 M urea washes. After de-crosslinking, enriched peptides were identified by tandem mass spectrometry analysis, and enrichment was quantified using dimethyl chemical labeling.

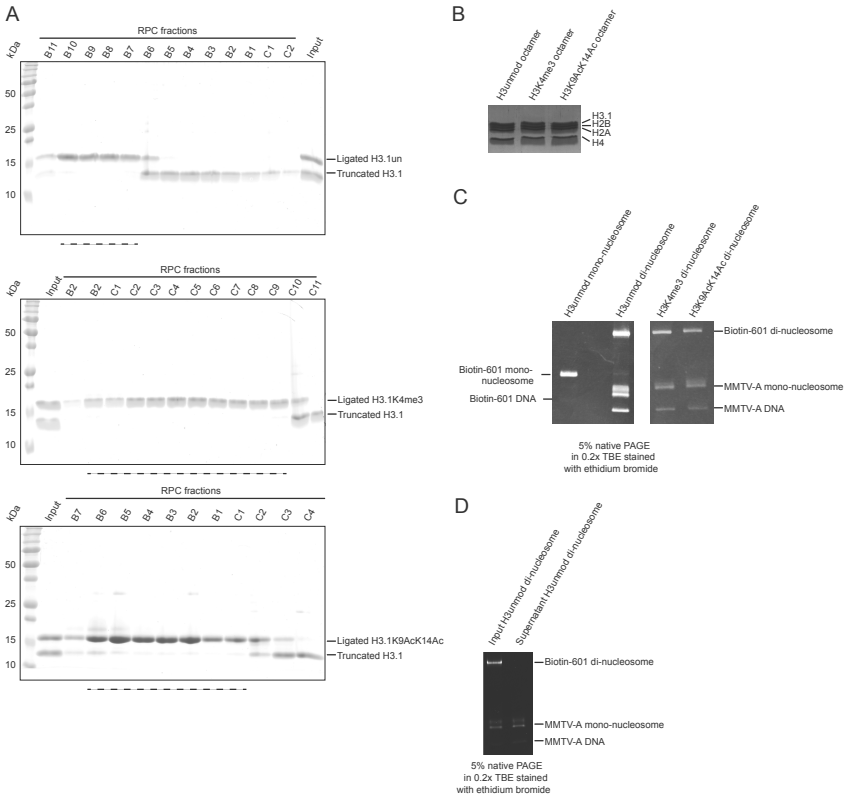
B Significantly enriched peptides for the XRCC5/6 heterodimer are colored on the crystal structure (PDB:1JEY) in red. Co-crystallized DNA is colored in blue.

C Outlier plot of enriched peptides for the mycG4 sequence. Each axis is the log2 transformed light/heavy dimethyl ratio from a single replicate. Each data point is a peptide. Enriched peptides using a significance cutoff of 1.5 interquartile ranges are colored in red. Background proteins are colored in black. Background proteins that are members of the SWI/SNF, PRC2, or NuRD complexes are colored in blue.

D Domain schematic of SMARCB1. The enriched peptide identified as significant in panel C is colored in red below the protein cartoon. Amino acid peptide indices are indicated by number.

E Binding curve for SMARCB1 (Fig. 2A, Cluster 8) to the mycG4 ssDNA oligo in NaCl binding conditions. The binding curve were generated by fitting the parameters of the Hill equation including *K*dApp. Each data point is the mean of three experiments (n=3), and the error bars represent the standard error of the mean.

F DNA pulldown and western blot experiments with bacterial lysates expressing either the SMARCB1 winged helix domain or the PHF10 double PHD finger. DNA pulldowns were performed as described as in the Methods using the mycG4 sequence and a variety of control sequences and analyzed by western blot. rs144361550 is a oligonucleotide representing a small insertion/deletion single nucleotide polymorphism predicted by computation to be G4-forming but characterized biochemically as non-G4 forming. The SP/KLF wild-type ssDNA is a GC-rich control sequence which is not predicted to form G4 structures. The SP/KLF mutated ssDNA is an AT-rich control sequence which is similarly not predicted to form G4 structures.



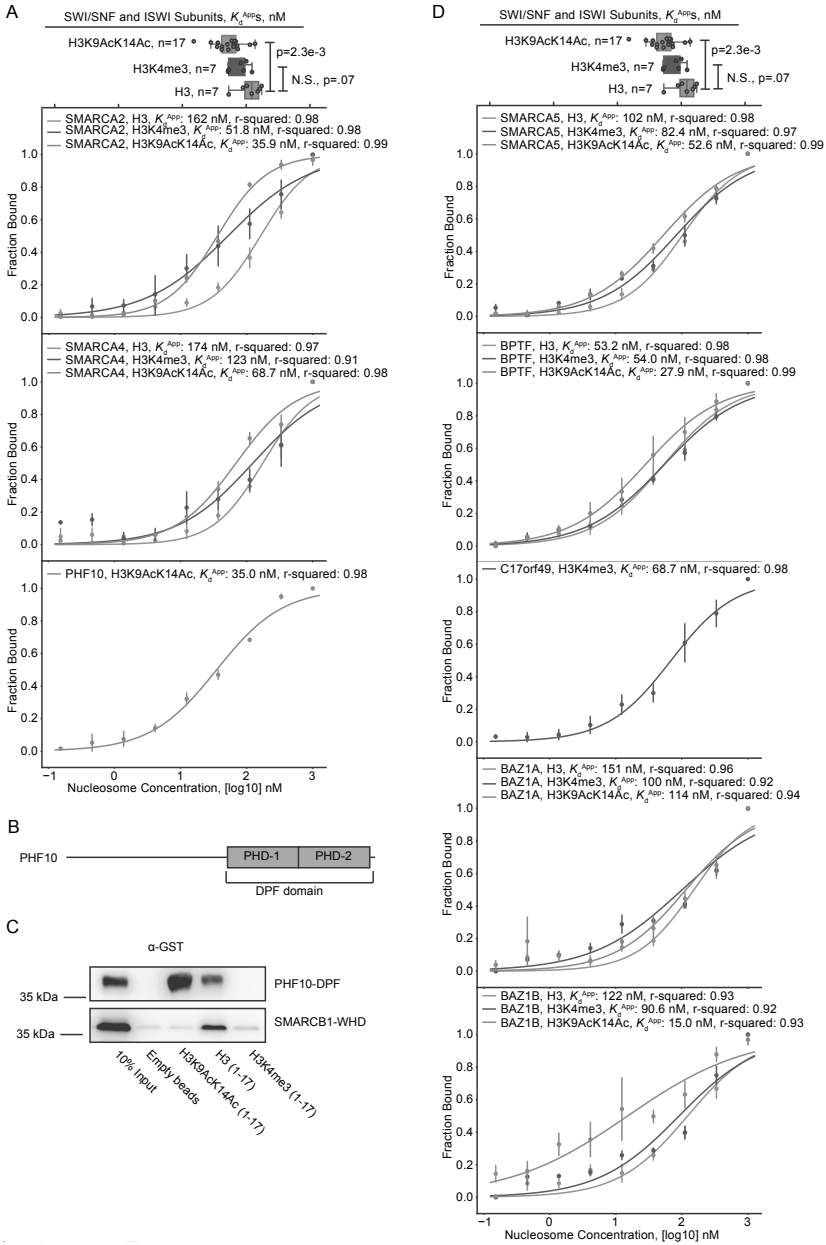
Supplementary Figure 9. Quality control checks of mono-nucleosomes and modified and unmodified di-nucleosomes

A 17.5% SDS-PAGE gels stained with Coomassie Brilliant Blue of Reverse-phase chromatography fractions indicating the separation of ligated full-length histone H3 (unmodified, K4me3 and K9AcK14Ac) and the truncated histone H3.1. Fractions underlined with a dashed line were pooled for refolding into histone octamers.

B 17.5% SDS-PAGE gels stained with Coomassie Brilliant Blue of the H3unmodified, H3K4me3 and H3K9AcK14Ac histone octamers.

C 5% native PAGE gels stained with ethidium bromide or SYBR safe of the biotinylated mono-/ di nucleosomes. The free DNA and band-shifted nucleosomes are indicated.

D Input di-nucleosome substrate and the supernatant after binding to streptavidin beads indicating that the free competitor MMTV-A DNA and mono-nucleosomes do not bind to the beads and the biotinylated di-nucleosomes do bind.



Supplementary Figure 10. SWI/SNF and ISWI di-nucleosome binding affinity is modulated by H3 modifications

A Binding curves for SMARCA2 (Fig. 3B, Cluster 5) and SMARCA4 (Cluster 5), SWI/SNF catalytic subunits, and PHF10 (Cluster 3), a SWI/SNF accessory subunit, are shown for unmodified, H3K4me3, and H3K9AcK14Ac modified di-nucleosomes. Above are boxplots of all identified *K*dApp values SWI/SNF and ISWI subunits for these substrates.

B Domain schematic of PHF10. The DPF domain used for cloning, protein expression, and DNA pulldown experiments is specifically indicated.

C Histone peptide pulldown and western blot experiments with bacterial lysates expressing either the PHF10 double PHD finger or the SMARCB1 winged helix. Peptide pulldowns were performed as described as in the Methods using the H3 peptides with the indicated post-translational modifications.

D Binding curves for SMARCA5 (Cluster 3), ISWI catalytic subunit, and BPTF (Cluster 3), C17orf49 (Cluster 1), BAZ1A (Cluster 4), and BAZ1B (Cluster 3), SWI/SNF accessory subunits, are shown for unmodified, H3K4me3, and H3K9AcK14Ac modified di-nucleosomes. Above are boxplots of all identified *K*dApp values SWI/SNF and ISWI subunits for these substrates. Significance is indicated based on a two sided t-test, with all measured values per sample, as indicated by the colored grouping, treated as sample populations.

Boxplots are displayed such that the center line is the median of the distribution, the box represents the first and third quartile of the data, and the whiskers represent 1.5 inter-quartile ranges. Significance is indicated based on a two-sided t-test, with all measured values per sample, as indicated by the colored grouping, treated as sample populations.

Binding curves were generated by fitting the parameters of the Hill equation including *K*dApp. Each data point is the mean of three experiments (n=3), and the error bars represent the standard error of the mean.

Supplementary Table 1. Oligonucleotides used in this study

Name	Sequence	Sequence 2
Sp/KLF Wild-type Consensus	/5Biosg/GAGAGCCCCG CCCCCTGGCT	AGCCAGGGGGCGGGG CTCTC
Sp/KLF Mutated	/5Biosg/GAGAGAAAAT AAAACTGGCT	AGCCAGTTTTATTTTCT CTC
AP-1	/5Biosg/AGTCGGCTAGC TGACTCAGGATGTCC	GGACATCCTGAGTC AGCTAGCCGACT
CTCF	/5Biosg/TCAGAGTGGC GGCCAGCAGGGGGCGC CCTTGCCAGA	TCTGGCAAGGGCGC CCCCTGCTGGCCGC CACTCTGA
E-box	/5Biosg/GGAAGCAGAC CACGTGGTCTGCTTCC	GGAAGCAGACCAC GTGGTCTGCTTCC
NF-Y	/5Biosg/ATTGACCAATC AGAGGTAGGATGAT	ATCATCCTACCTCTG ATTGGTCAAT
TATA-box	/5Biosg/GCGGCGCTCTA TATAAGTGGGCAGTG	CACTGCCCACTTATA TAGAGCGCCGC
TEAD	/5Biosg/TCGGGACCCAGG CCTGGAATGTTTCCACC	GGTGGAAACATTCC AGGCCTGGGTCCCGA
MycG4	/5Biosg/TGGGGAGGGTGG GGAGGGTGGGGAAGG	
Telomere	/5Biosg/TTAGGGTTAGG GTTAGGGTTAGGG	
rs144361550_PARP1_ins_biotin_ rev (predicted G4, biochemically characterized as non-G4)	/5Biosg/GAGCGAGCGGGCCCGGGCCCgggcccT CGGAGCGGCACTTGGGGCC	

*Free sequences for competition experiments were ordered without biotinylation

Supplementary Table 2 Antibodies and recombinant proteins used in this study

Factor	Source	Catalog Number	Ab Dilution Used
SP1	Sigma	S9809	1:100
SP3	abcam	ab72594	1:500
KLF4	Sigma	HPA002926	1:1000
EZH2	Cell signaling technologies	5246	1:500
GATAD2B	Bethyl laboratories	A301-282A	1:2000
Suz12	abcam	ab12073	1:2000
GST	Thermo	MA4-004	1:1000
Recombinant protein used			
Protein	Source		
SP3	Abnova	H00006670-P01	
KLF4-ZF	Spruijt et al. Cell. 2012.		
PHF10 Double PHD finger domain (aa358-aa498)	This study	For primer: CCCGGATCCCCAAACG TTCTGTACTGTCC	Rev Primer: GGGAAGCTTATTAT CCCTCTTTGCTGTT TTTC
SMARCB1 winged helix domain (aa1-aa112)	This study	For primer: CAATCCATGGGAATG ATGATGATGGCGCT GAG	Rev Primer: TCGGGATCCTTATTA GATGGACACAGCCTT GTAC

Supplementary Table 3 Public data sets used in this study

Factor	Cell Line	Database	Accession Number
ChIP-Sequencing			
SMARCB1	HeLaS3	ENCODE	ENCFF002CSN
SMARCB2	K562	ENCODE	ENCFF993YKH
GATAD2B	MCF7	ENCODE	ENCFF046BRP
EED	GM12878	ENCODE	ENCFF023ALY
EZH2	NHEK Keratinocytes	ENCODE	ENCFF002CFB
EZH2	GM12878	ENCODE	ENCFF615NYO
EZH2	Primary human CD20+ B-cells, RO01794	ENCODE	ENCFF434OEY
G4-sequencing			
Induced G4 (K+, PDS), plus strand	Primary human B-cells, NA18507	GEO	GSE63874
Induced G4 (K+, PDS), minus strand	Primary human B-cells, NA18507	GEO	GSE63874

Chapter 5

Discussion

[They go] into non-being, which is to say, everything.
-*Harry Potter and the Deathly Hallows*, J.K. Rowling

The emergence of cancer genomics as a major discipline in biomedical research has precipitated the discovery of a genetic menagerie of disease related genotype-phenotype associations and putative functional sequence variants. The benefit of hypothesis-generating, data-driven, multi-omics science is that we now have far more molecular and genetic information about the cellular state that is "cancer" than ever before. But the ability to describe the features of a certain state does not equal a functional, mechanistic understanding of how that state arose, how it behaves, or what causes it to respond in one way or another. In the 18th century, Carl Linnaeus attempted to describe and categorize all living things according to their features, similarities, and differences. But a taxonomy is a *descriptive* categorization of life. Not until the insights of Darwin and Wallace, and the concept of common descent, variation, and natural selection, did a satisfying mechanistic explanation for the diversity of living things arise. The challenge currently facing the (biomedical) molecular biology community is how to translate the continually increasing deluge of *information* into *knowledge*. In other words, how to turn the *molecular taxonomy* of genetic variants and molecular phenotypes that is currently being produced into a satisfying mechanistic explanation for how a cell, tissue, organ, organism, genome, proteome, epigenome, metabolome, protein, protein complex, etc. *actually behaves*, or *misbehaves* in the context of disease? How to turn a *description* into a *prediction*?

This thesis has presented positive cases where mass spectrometry was leveraged to connect a particular DNA sequence, variant, or nucleosome with a set of specific or high affinity interactors; however, these examples all exhibit one serious disadvantage. It is the very same disadvantage that, at the moment, all protein-based studies possess compared with DNA, RNA, or sequencing based studies. All research presented here was conducted essentially on a "case-by-case" basis, which in essence means it was slow compared to high-throughput sequencing approaches. Sequencing data is being produced at a pace that rapidly outstrips the speed with which comparable proteomics data can currently be produced. Many more mutations and sequence variants are being discovered via NGS and whole-genome sequencing (WGS) than can reasonably be profiled by proteomics methodologies, or by any other standard molecular biological or biochemical technique for that matter. Thus, it is critical that future research incorporates innovative high-throughput methods for profiling

many (i.e. hundreds or even thousands) of sequence variants. In Chapter 4 of this thesis, we use a filter-plate based system which can in principle perform ~100 affinity purifications in parallel. However, this approach still only allows for semi-quantitative profiling of ~25 variants per 96-well plate or absolutely quantitative profiling of 8 sequences or baits. Maximizing the number of sequence variants that can be profiled with filter plate based systems will likely require robotic automation. Similarly, microfluidics platforms could enable automated and standardized affinity purification protocols while having the additional benefit of dramatically downscaling the material requirements (i.e., protein and DNA amounts) per affinity purification. On the other hand, pooling and deconvolution experimental designs might be a promising experimental avenue for increasing the throughput of variant profiling by MS without demanding additional technology investment. In any case, though there have recently been a few isolated cases of research studies reporting 1000+ affinity purifications in a single manuscript, these cases are still few and far between. Future developments in the automation of DNA focused AP-MS experiments and analysis will undoubtedly be a subject of keen interest in a variety of multi-disciplinary fields.

The canonical role of transcription factors in binding to specific DNA sequence motifs, often monomerically or in limited size multimers such as dimers or trimers, and regulating transcription of target genes via direct interactions with the core transcriptional machinery or chromatin remodeling or modifying enzymes is by now relatively well established. Indeed, in Chapter 2 we reported mass spectrometry studies of precisely this type of canonical DNA binding and gene regulatory function. On the other hand, our analysis of tetrameric GABP binding at novel and endogenous *TERT* promoter mutations already suggests that more than simple motif creation or disruption can influence TF binding. In that case, it was the precise *spacing* of nearby ETS motifs that dictated stable GABP binding. In contrast, this motif spacing appeared unimportant for GABP binding at *SDHD* promoter mutations, showing that motif architecture itself can be a context-specific form of regulation. Indeed, understanding the role that motif spacing plays in facilitating co-regulation of target genes via direct or functional interactions between TFs binding at co-occurring motifs is an important topic of current research and an area where mass spectrometry protein-DNA interactomics is sure to play a role. However,

we observe that in addition to motif spacing as a higher-order determinant of TF binding, *DNA structure* can itself act as recognition site for some proteins and protein complexes. Although such *structure-specific* interactions have been reported previously, most often for DNA helicases, we report in Chapter 3 that *differential* DNA structures can also regulate transcription allele-specifically for cancer-associated insertion/deletion variants. Future research might well consider if, and to what extent, small insertion/deletion variants in repetitive regions (i.e., variants that add no novel sequence or motif content) regulate transcription via structure-specific interaction normally and abnormally if the variant is associated with cancer risk. Additionally, we show in Chapter 4 that structure-specific DNA recognition may be more widespread than previously realized, given that ligand affinity is marginally but not dramatically decreased when DNA structure is chemically disrupted. Such structure-specific recognition might not be readily identified using traditional AP-MS outlier calling methods. G-quadruplex RNA recognition has been reported for PRC2, yet there is very little in the literature connecting G-quadruplexes or a litany of other DNA secondary structures with proteins and protein complexes that regulate chromatin state. That G-quadruplexes are preferentially observed in regulatory regions that are nucleosome depleted suggests at least a connection between DNA structural elements and chromatin maintenance. To what extent this connection exists, and its functional implications, will surely be a topic for future research in chromatin biochemistry and structure. In addition, profiling the high affinity "readers" of the various reported non-G4 DNA secondary and higher-order structures (including i-motifs, cruciform, triplexes, alternate quadruplexes including inter-molecular quadruplexes, and hairpins) would shed some much needed light into regulation of, and regulation by, these interesting and largely uncharacterized structural elements.

In Chapter 4 of this thesis, we identify high affinity interactors of (modified) nucleosomes and predominantly di-nucleosomes. Although we did not expect the "compactness" of di-nucleosomes (as opposed to, for example, 12mer or 24mer nucleosome arrays) to make a major contribution to the affinities we measured, it is well known that some histone PTMs including histone acetylation can dramatically affect the stability, DNA wrapping, and high-order compaction of nucleosomes and nucleosome arrays. Therefore, we cannot exclude the possibility that nucleosome or DNA accessibility might increase

or decrease protein-nucleosome binding affinities differentially based on the nucleosome modification state. Surprisingly, however, the binding assay we describe in Chapter 4, along with differentially modified nucleosomes and nucleosome arrays of different lengths, actually enables new experimental strategies for studying the effect histone PTM mediated nucleosome compaction has on protein-nucleosome binding affinities. In addition to measuring protein-nucleosome affinities for nucleosome arrays of the same length but with different PTMs, one could in principle measure nucleosome arrays of different lengths but with the same set of PTMs. In doing so, if different affinities were measured for nucleosome arrays of different lengths, it could be inferred that any changes in affinity would be due to the structural factors of nucleosome compaction, accessibility, and stability. Additionally, analysis of the rate of affinity changes might provide additional structural/functional biochemical data: a rapidly saturating decrease in affinity with longer nucleosome array length could be correlated with maximally compact units of few nucleosomes, for example. It has been well known for some time that chromatin is a plastic and dynamic molecular structure; an analysis of the PTM-dependent relationship between protein complex affinity and nucleosome/chromatin compaction would add to our knowledge of the dynamic structural properties of chromatin. Finally, most studies with recombinant nucleosome systems use a relatively small number of well-defined sequences (synthetic nucleosome positioning sequences, palindromic alpha satellite repeats, etc). A major next step is to use different, biologically interesting sequences to assess how underlying DNA *sequences* interact with histones and histone PTMs to influence compaction (or indeed de-compaction) of chromatin via specific TF or chromatin factor interactions.

What are the next imminent advances in the field of mass spectrometry based protein-DNA interactomics? Cross-linking mass spectrometry (XL-MS) has advanced rapidly in the past decade or so, and now XL-MS studies proteome-wide identifying thousands of unique residue-level cross-linked peptide pairs are seen somewhat regularly. RNA XL-MS studies using UV cross-linking have been a more recent research topic. Presumably, some of the advances made in protein and RNA XL-MS technologies will begin finding applications in protein-DNA cross-linking studies, potentially in combination with hydrogen-deutrium exchange or limited proteolysis approaches. The ability to identify direct DNA binding proteins from a DNA affinity purification experiment, at a

residue level, will be broadly useful. Additionally, Native MS and Ion Mobility MS are promising methods for studying structural and biochemical aspects of protein-DNA interactions in near native conditions, including stoichiometry information and conformational states or changes. Ensemble, hybrid methods that combine some or all of these data types with additional structural data, such as SAXS or low resolution EM data, may be more broadly accessible as data collection, integration, and modeling technologies become optimized and generalized. Finally, there has been some long-term interest in developing pharmaceutical ligands to specifically inhibit sequence-specific DNA binding properties of transcription factors. Though these efforts are still very much ongoing, DNA-protein AP-MS approaches will likely be instrumental in profiling the potency and specificity of future TF inhibitors, possibly in combination with Thermal Proteome Profiling experiments to assess off-target effects.

Since Jacob and Monod's famous work in the 1950's and 1960's, it has become increasingly clear that the early model of a single transcription factor regulating a single gene that becomes a single protein is little more than fantasy. We now know that each transcription factor binds in many places in the genome, and in addition, these locations often differ depending on the type of cell in question. The functions enacted by TFs are complex, involving transcriptional activation and repression, recruitment of the basal transcriptional machinery, regulating chromatin state, and mediating genome folding. Furthermore, this complexity increases exponentially given the interplay between different TFs and between their various effector functions. As such, the dysregulation of TF mediated transcriptional regulatory processes in disease states can be similarly complex. It is very true that, if you love complex systems, now is a golden time to be a molecular biologist. But there is much more to be gained from integrative science that combines a top-down, systems-level approach to descriptive biology with a bottom-up, mechanistic approach to defining molecular functions. Understanding a cell, understanding the sum of the parts, still requires a reductionist, bottom-up understanding of the parts. And it is here that mass spectrometry based protein-DNA interaction proteomics has made, and will continue to make, considerable scientific contributions in biology.

Chapter 6

Summary / Samenvatting

Summary

Throughout this thesis, I have argued by demonstration that mass spectrometry is a useful tool for connecting a description (a variant, a genetic association, a sequence) with a prediction (a regulatory TF or protein). I first introduce in **Chapter 1** the scientific background of this thesis, giving a thorough overview of transcriptional gene regulation and chromatin biology, cancer genomics, and mass spectrometry. I emphasize in particular the opportunities for collaboration between these traditionally somewhat delimited fields.

In **Chapter 2**, we identified a TF, GABP, that binds sequence-specifically to novel ETS motifs formed by recurrent *TERT* promoter mutations in melanoma and other cancers, yet is inhibited from binding by the disruption of endogenous ETS motifs by *SDHD* promoter mutations. Particularly in the case of the *TERT* promoter, the genetic disease association of the recurrent mutations was described relatively far before a functional mechanism was proposed. Yet the observation that GABP binds sequence-specifically at the mutation-induced ETS motifs immediately suggested just such a functional mechanism: that GABP binds mutation specifically and, in the manner of many TFs, activates transcription of *TERT*. The identification of mutation-specific binding at *SDHD* followed immediately from a search for mutation-induced motif changes with a similar genetic signature.

Chapter 3 describes a biologically more complex case with a similar conceptual rationale. In contrast to the *TERT* and *SDHD* promoter mutations, the putative functional germline variant at the *PARP1* melanoma locus was an insertion/deletion SNP that fell in a hexameric (GGGCCC) repeat that created no new sequence motifs. As such, it was highly unlikely that any putative protein regulator binding at that site would be a canonical sequence-specific TF. Instead, we identified a number of proteins annotated as binding preferentially to DNA structural elements including G-quadruplexes. Further analysis showed an insertion allele preferential binding protein, the DNA helicase RECQL, transcriptionally regulated *PARP1* expression. While the insertion allele was predicted by computation to form a G-quadruplex and the deletion allele was not, we observed that any differential secondary structures between the insertion and deletion alleles was not due to canonical G-quadruplex formation, highlighting the diversity of possible DNA secondary structures. Yet, intriguingly, additional proteomics analysis suggested that ligand stabilized DNA secondary structures

dramatically affected protein binding. In any case, the underlying logic was the same; a DNA structural variant was connected, by mass spectrometry, with a putative functional regulator to offer a mechanistic explanation for a disease association.

Furthermore, **Chapter 4** shows that mass spectrometry can be utilized to offer biochemical *affinity* information about protein-DNA interactions in addition to the semi-quantitative specificity data described in the previous chapters. It is relevant to note, at this point, that not all protein-DNA regulatory interactions are completely equal. On the contrary, they differ in both the specificity and the affinity of the protein-DNA pair in question. We show in Chapter 4 that many putative regulatory TFs can bind the same DNA sequence specifically, but with quite different absolute affinities. Furthermore, in addition to estimating absolute affinities, we show how chemical modulation of G-quadruplex structure can decrease high affinity binding of some chromatin remodeling and modifying enzymes. We show a similar modulated binding affinity for many proteins and protein complexes by histone tail PTMs in the context of nucleosomes, and in fact, we show that high affinity binding may be modulated by histone PTMs for only some, but not all, subunits of large chromatin remodeling complexes. But more generally, what this approach facilitates is a "narrowing down" of putative functional regulators, under the general hypothesis that higher affinity interactors are more likely to be functional than lower affinity interactors.

Finally, **Chapter 5** provides a more colloquial discussion of the scientific contents of the thesis. I generally contextualize the contents of this thesis, focusing on limitations of the described research and research methods while also emphasizing opportunities and directions for future research.

In general, this thesis has contributed to our understanding of sequence-specific transcription factor-DNA interactions and their deregulation in cancer. Concurrently, this thesis has established mass spectrometry based interaction proteomics as a powerful tool for identifying such sequence-specific protein-DNA interactions and has developed a new mass spectrometry assay for quantifying these interactions.

Samenvatting

In dit proefschrift heb ik laten zien dat massaspectrometrie een geschikt hulpmiddel is om een feitelijkheid (een variant, een genetische associatie, een sequentie) te verbinden met een voorspelling (een regulerende transcriptie factor (TF) of eiwit). In **Hoofdstuk 1** introduceer ik eerst de wetenschappelijke achtergrond van dit proefschrift, met een grondig overzicht van transcriptionele genregulatie en chromatine biologie, kankergenetica, en massaspectrometrie. Ik benadruk met name de mogelijkheden voor samenwerking tussen deze twee – traditioneel enigszins afgebakende – onderzoeksgebieden.

In **Hoofdstuk 2** hebben we een TF geïdentificeerd, GABP, die sequentie-specifiek bindt aan een nieuw ETS-motief dat gevormd wordt door herhaaldelijk voorkomende mutaties in de *TERT*-promotor in melanoom en andere vormen van kanker, maar waarvan de binding wordt verhinderd door het verstoren van endogene ETS-motieven door mutaties in de *SDHD*-promotor. In het bijzonder in het geval van de *TERT*-promotor was het genetische ziekteverband met de herhaaldelijk voorkomende mutaties al relatief lang beschreven voordat hiervoor een functioneel mechanisme werd voorgesteld. Toch suggereert de observatie dat GABP sequentie-specifiek bindt aan de ETS motieven die zijn ontstaan door mutaties meteen een dergelijk functioneel mechanisme: dat GABP specifiek aan de mutatie bindt en, zoals veel TFs, transcriptie van *TERT* activeert. De identificatie van mutatie-specifieke binding aan *SDHD* volgde direct uit een zoektocht naar veranderingen van motieven door mutaties met soortgelijke genetische kenmerken.

Hoofdstuk 3 beschrijft een biologisch complexer geval met een soortgelijke conceptuele redenering. In tegenstelling tot de mutaties op de *TERT* en *SDHD* promotors was de vermeende functionele kiembaanvariant op het *PARP1* melanoom locus een insertie/deletie SNP, gelegen in een hexamere repetitieve regio (GGGCCC), die geen nieuwe sequentiemotieven creëerde. Daardoor was het hoogst onwaarschijnlijk dat een eventuele vermeende eiwitregulator die op die plek kon binden een canonieke sequentie-specifieke TF zou zijn. In plaats daarvan hebben we een aantal eiwitten geïdentificeerd waarvan bekend was dat die preferentieel binden aan structurele DNA-elementen zoals G-quadruplexen. Verdere analyse wees uit dat de DNA helicase RECQL, die preferentieel bindt aan het insertieallel, de expressie van *PARP1* transcriptioneel reguleerde. Hoewel computeranalyses voorspelden dat het insertieallel een G-quadruplex

zou vormen en het deletieallel niet, constateerden we dat eventuele differentiële secundaire structuren tussen de insertie- en deletieallelen niet te wijten waren aan canonieke G-quadruplex formatie, wat de diversiteit van mogelijke DNA secundaire structuren benadrukt. Intrigerend was daarom dat verdere proteomicsanalyse suggereerde dat eiwitbinding dramatisch werd beïnvloed door ligand-gestabiliseerde DNA structuren. In elk geval was de onderliggende logica hetzelfde; een structurele DNA variant werd door massaspectrometrie gelinkt aan een vermeende functionele regulator, waarmee een mechanistische verklaring voor een ziekteverband werd gegeven.

Daarnaast laat **Hoofdstuk 4** zien dat massaspectrometrie gebruikt kan worden om biochemische informatie over de affiniteit van eiwit-DNA interacties te verkrijgen, naast de semi-kwantitatieve specificiteitsgegevens zoals beschreven in de voorgaande hoofdstukken. Op dit punt is het relevant om op te merken dat niet alle regulerende eiwit-DNA interacties volledig gelijkwaardig zijn. In tegendeel, ze verschillen in zowel de specificiteit als de affiniteit van het betreffende eiwit-DNA-koppel. We laten in hoofdstuk 4 zien dat veel vermeende regulerende TFs specifiek aan dezelfde DNA sequentie kunnen binden, maar met vrij verschillende absolute affiniteiten. Naast het schatten van absolute affiniteiten laten we bovendien zien hoe de chemische modulatie van de G-quadruplex structuur de hoge affiniteitsbinding van sommige chromatine remodellerende en modificerende enzymen kan veranderen. We laten voor veel eiwitten en eiwitcomplexen een bindingsaffiniteit zien die soortgelijk gemoduleerde wordt door post-translationele modificaties (PTMs) op histonstaarten in de context van nucleosomen. In feite laten we zien dat alleen voor sommige, maar niet alle, subeenheden van grote chromatine remodellerende complexen de hoge affiniteitsbinding gemoduleerd kan worden door histon PTMs. Meer in het algemeen, wat deze aanpak mogelijk maakt is het selecteren van vermeende functionele regulatoren, onder de algemene hypothese dat het meer waarschijnlijk is dat interactoren met een hogere affiniteit functioneel zijn dan interactoren met een lagere affiniteit.

Ten slotte geeft **Hoofdstuk 5** een meer informele discussie over de wetenschappelijke inhoud van het proefschrift. Ik plaats de inhoud van het proefschrift in een meer algemene context, waarbij ik focus op de beperkingen van het beschreven onderzoek en onderzoeksmethoden en tegelijkertijd ook de mogelijkheden en richtingen voor vervolgonderzoek benadruk.

Dit proefschrift heeft algemeen bijgedragen aan ons begrip van sequentie-specifieke interacties tussen TFs en DNA en hun deregulering in kanker. Tegelijkertijd heeft dit proefschrift op massaspectrometrie gebaseerde interactie-proteomics vastgesteld als een krachtig hulpmiddel om zulke sequentie-specifieke eiwit-DNA interacties te identificeren en beschrijft het een nieuwe massaspectrometrie methode, ontwikkeld voor het kwantificeren van deze interacties.

[Translation into Dutch by Ino Karemaker, to whom I offer my sincere gratitude.]

Chapter 7

Acknowledgments

It seems obvious to me that this thesis and the research accompanying it would not have been even remotely possible without the training, support, encouragement, and sacrifices of a number of incredible and noteworthy people. I hope I have made it clear to these people along the way how important and significant their presence in my life has been. Despite my best and most sincere efforts, I'm sure I will omit some names via my own fault and in some attempt to maintain brevity as a complete acknowledgment would require the writing of another book and more. Any errors of omission should not be taken as a sign of anything other than my own disorganized nature, which I'm sure, if you've been with me this far, you've already forgiven me for many times along the way.

So, I can only say, very sincerely, thank you all. That I have shared this with all of you means more to me, certainly, than I can say or write.

And in particular, I would like to say thank you:

To my mother and father, always. My mother was my first teacher, and my father was my first coach. You taught me, before anyone else, how to learn and how to compete and set goals. I love you and owe you both more than I can ever say.

To my brother and my sister. In so many ways, we couldn't be much more different, and yet we couldn't be any closer. Being your older brother and being in our family has made me who I am, even more than my academic training. And you all, the whole family, keep me humble, which is something I think you take a great amount of pride in.

To my Grandpa. Maybe more than anyone else in our family, I think we are the same. More than anyone else in the family, and more than me, you embody the mind and spirit of a PhD, which I will always believe you deserve to have. You are knowledgeable, curious, and wise, and I will always value the meandering conversations we've had and everything I've learned from them.

To Michiel. Your lab was a home to me for almost five years. I understand now how you took a risk on me, but it was a great pleasure to grow under your mentorship into the scientist and academic that five years ago you saw, perhaps,

the potential for me become. I am a trained scientist and, more importantly, a better person for having learned and worked in the lab you built.

To David, Susan, Josh, Kevin, and Jiyeon. Before ever I started a PhD, you took a very raw and naïve young student and gave him a chance to learn how knowledge is made. You were the first to give me the possibility and the direction to start down this path.

To Susan, Raghu, Xiaofei, Cristina, Nelleke, Vicky, and Guido. From watching all of you, I learned how science should be done. You taught me directly and by example, and I couldn't have imagined better examples.

To Irem, Arne, Ino, Lisa, Rik, and Jelmer. All of you I learned from, even as we were going through this process together. All of you I will learn from in the future, as I follow your careers and your future research.

To Pascal and Marijke. It's clear the lab would collapse into chaos without you. Thank you for teaching me, correcting me, and working with me, but also for keeping everything together.

To Esther, Ian, Tabea, Cathrin, and Laura. I am entirely certain I learned more from supervising and working with you than you ever learned from me. One of the most satisfying and valuable aspects of my PhD was watching and helping you all develop into great and talented scientists.

To the other students in the Vermeulen Group, including more recently Jan, Freek, Agon, Irene, Hannah, and Simone. In the same way, our interactions and the process of watching all of you develop into bright and capable scientists has been a great pleasure.

To Henk. The department in so many ways exists because of your influence. In the same way, thank you for giving all of us the possibility of doing the science we love.

To Gert Jan. For envisioning what the DevCom project could be, for making it real, and for giving all of us a place to grow and learn.

To all of my colleagues in the department, in the SuperGroup, at Utrecht, at the Hubrecht, and in the DevCom network. Half of doing science is sharing it with

people you enjoy and are enriched by. I was challenged by you, learned from you, celebrated with you, and struggled with you. In the end, happiness is only real when shared, and I was happy to share it with all of you.

To my scientific collaborators. I have appreciated in equal parts the scientific contributions you have made to my projects, and the experience I gained from contributing to your projects.

To the Roberts family. You are all kind and good, and I cannot thank your enough for welcoming my family and I into your lives. I learned from you, beyond science and academia, that sometimes it's just the right move to make a risky four bid and lead with a seven.

Finally, completely, to Maddie, who stuck with me to the very end. It's not the book I started for you, but at some point it became the book I was finishing and did finish for you. I look forward more than I can say to whatever it is that we'll become.

Chapter 8

Curriculum vitae

Curriculum vitae

Matthew Michael Makowski graduated secondary school from Faith Christian Academy in Sellersville, Pennsylvania, the United States of America. He earned his B.S. in Biology with Honors, *Summa Cum Laude*, with a Minor Degree in Literature, from American University, Washington, D.C., the United States of America. At American University, he performed research on evolutionary genetics submitted as an Honors Thesis under the supervision of Dr. David Carlini. Concurrently, he performed a summer internship working on the structural biology of viral capsids in the lab of Dr. Susan Hafenstein under the direct supervision of Dr. Josh Yoder at the Pennsylvania State University College of Medicine, Hershey, Pennsylvania, the United States of America. After completing his undergraduate studies, he worked on the molecular cancer genomics of melanoma, focusing on post-GWAS functional analysis of disease-associated SNPs, in the lab of Dr. Kevin Brown under the direct supervision of Dr. Jiyeon Choi at the Laboratory of Translational Genomics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, the United States of America. He completed his graduate studies with a PhD in Molecular Biology working on protein-DNA interactions using mass spectrometry in the lab of prof. dr. Michiel Vermeulen at Radboud University Nijmegen, Nijmegen, the Netherlands. During that period, he was a Marie Curie Fellow in the DevCom Marie Curie ITN led by prof. dr. Gert Jan Veenstra. He will next perform a Post-Doctoral Fellowship in the group of Dr. Karolin Luger working on chromatin structural biology and biochemistry at the University of Colorado in Boulder, Colorado, the United States of America.

Publications

Publications included in this thesis

1. **Makowski MM***, Willems E*, Fang J*, Choi J, Zhang T, Jansen PW, Brown KM*, Vermeulen M*. An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations. *Proteomics*. **2016**.
2. Zhang T*, Xu M*, **Makowski MM***, Lee C, Kovacs M, Fang J, Willems E, Trent JM, Hayward NK, Vermeulen M*, Brown KM*. SDHD Promoter Mutations Ablate GABP Transcription Factor Binding in Melanoma. Cancer Research. 2017.
3. Choi J*, Xu M*, **Makowski MM**, Zhang T, Law MH, Kovacs MA, Granzhan A, Kim WJ, Parikh H, Gartside M, Trent JM, Teulade-Fichou MP, Iles MM, Newton-Bishop JA, Bishop DT, MacGregor S, Hayward NK, Vermeulen M, Brown KM. A common intronic variant of PARP1 confers melanoma risk and mediates melanocyte growth and senescence escape via regulation of MITF. *Nature Genetics*. **2017**.
4. **Makowski MM**, Gräwe C*, Nguyen NV*, Foster B*, Bartke T#, Vermeulen M#. Proteome-wide affinity quantification by mass spectrometry. *Nature Communications.* 2018.

Publications not included in this thesis

1. Robles-Espinoza CD, Harland M, Ramsay AJ, Aoude LG, Quesada V, Pritchard AL, Tiffen JC, Petljak M, Palmer JM, Symmons J, Johansson P, Stark MS, Gartside MG, Snowden H, Montgomery GW, Martin NG, Liu JZ, Choi J, **Makowski MM**, Brown KM, Keane TM, López-Otín C, Gruis NA, Hayward NK, Bishop T, Newton-Bishop JA, Adams DJ. *POT1* Mutations predispose to familial melanoma. *Nature Genetics*. **2014**.
2. Kloet SL*, Baymaz HI*, **Makowski M***, Groenewold V, Jansen PW, Berendsen M, Niazi H, Kops GJ, Vermeulen M. Towards elucidating the stability, dynamics and architecture of the nucleosome remodeling and deacetylase complex by using quantitative interaction proteomics. *The FEBS Journal*. **2015**.
3. Carlini DB, **Makowski M**. Codon bias and gene ontology in holometabolous and hemimetabolous insects. *Journal of Experimental Zoology. Part B, molecular and developmental evolution*. **2015**.

4. **Makowski MM**, Willems E, Jansen PW, Vermeulen M. Cross-linking immunoprecipitation-MS (xIP-MS): Topological Analysis of Chromatin-associated Protein Complexes Using Single Affinity Purification. *Molecular & Cellular Proteomics*. **2016**.
5. Kloet SL, **Makowski MM***, Baymaz HI*, van Voorthuijsen L, Karemaker ID, Santanach A, Jansen PW, Di Croce L, Vermeulen M. The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nature Structural & Molecular Biology*. **2016**.
6. Zhang X*, Smits AH*, van Tilburg GB*, Jansen PW, **Makowski MM**, Ovaa H*, Vermeulen M*. An Interaction Landscape of Ubiquitin Signaling. *Molecular Cell*. **2017**.
7. Fang J*, Jia J*, **Makowski M**, Xu M, Wang Z, Zhang T, Hoskins JW, Choi J, Han Y, Zhang M, Thomas J, Kovacs M, Collins I, Dzyadyk M, Thompson A, O'Neill M, Das S, Lan Q, Koster R; PanScan Consortium; TRICL Consortium; GenoMEL Consortium, Stolzenberg-Solomon RS, Kraft P, Wolpin BM, Jansen PWTC, Olson S, McGlynn KA, Kanetsky PA, Chatterjee N, Barrett JH, Dunning AM, Taylor JC, Newton-Bishop JA, Bishop DT, Andresson T, Petersen GM, Amos CI, Iles MM, Nathanson KL, Landi MT, Vermeulen M, Brown KM*, Amundadottir LT*. Functional characterization of a multi-cancer risk locus on chr5p15.33 reveals regulation of TERT by ZNF148. *Nature Communications*. **2017**.
8. Pines A*, Dijk M*, **Makowski M**, Meulenbroek EM, Vrouwe MG, Mullenders LH, Vermeulen M, Vermeulen W, van Attikum H. TRiC chaperonin controls transcription resumption after UV damage by regulating Cockayne Syndrome protein A. *Nature Communications.* 2018.

Radboud University